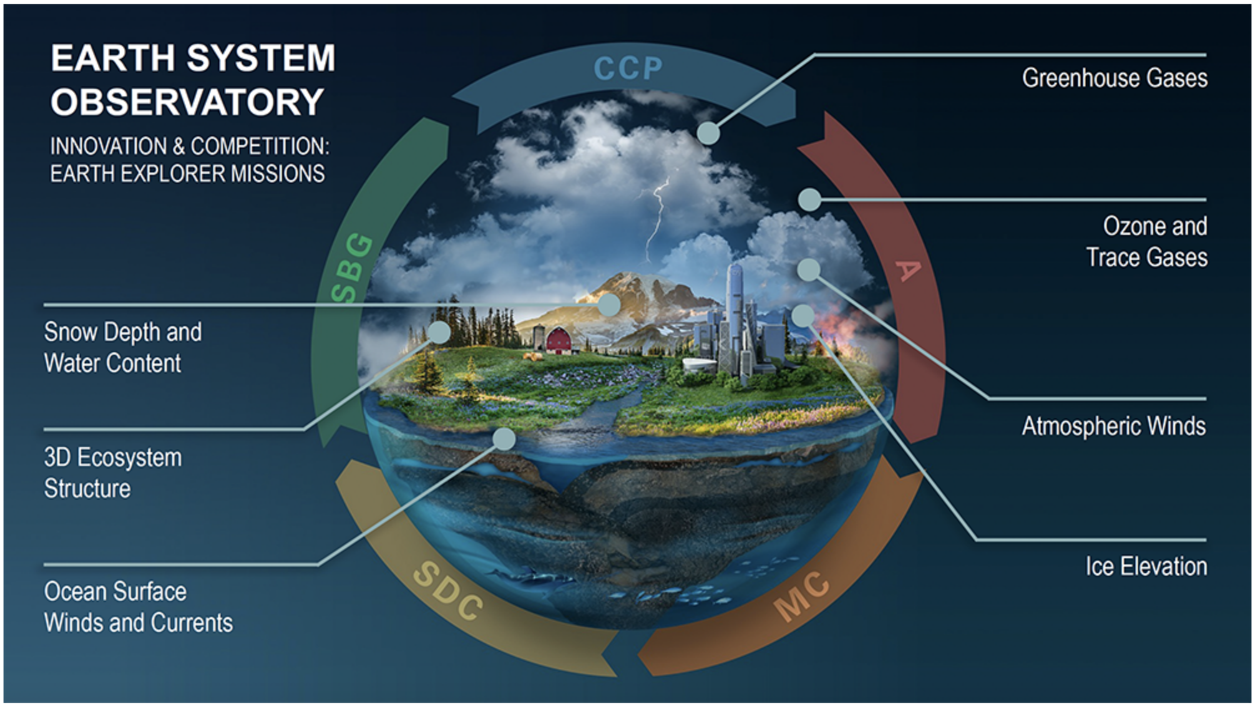
Workshop #2 Report

ESO Mission Data Processing Study: Summary of State-of-the-Practice and State-of-the-Art Mission Data Processing System Architectures



[A portion of] this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004). Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

Contributors

E. Natasha Stavros (Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder)

Elias Sayfi (Jet Propulsion Laboratory (JPL), California Institute of Technology)

Andrew Michaelis (NASA Ames Research Center)

Bernie Bienstock (Jet Propulsion Laboratory (JPL), California Institute of Technology)

Wenying Su (NASA Langley Research Center)

Hook Hua (Jet Propulsion Laboratory (JPL), California Institute of Technology)

Evelyn Ho (NASA Goddard Space Flight Center)

Karen Yuen (Jet Propulsion Laboratory (JPL), California Institute of Technology)

Qing Yue (Jet Propulsion Laboratory (JPL), California Institute of Technology)

Curt Tilmes (NASA Goddard Space Flight Center)

Lesley Ott (NASA Goddard Space Flight Center)

Chris Engebretson (USGS)

Adrian Parker (NOAA/NESDIS)

Sean Harkins (Marshall Space Flight Center)

Mike Chepurin (NOAA/NESDIS)

Table of Contents

1. Purpose and Scope

The ***purpose*** of this report is to synthesize findings from the NASA Earth System Observatory (ESO; Margetta, 2021) Processing Workshop #2. The full agenda and announcement is found here: https://earthdata.nasa.gov/esds/open-science/oss-for-eso-workshops.

This workshop is the second of three in the "Open Source Science For ESO Mission Data Processing Study". Workshop #1 focused on gathering needs and considerations for evaluating different open science data system architectures to support Earth system science and mission data system efficiencies. The findings from that workshop along with the criteria for evaluating future Mission Data Processing System (MDPS) architectures against the study objectives (Section "ESO Study Overview") can be found at: http://hdl.handle.net/2014/53042. The goal of workshop #2 was to understand the state of the practice and state-of-the-art in Big Data processing systems.

This document acts as a documentation of key points made during the workshop by study stakeholders, and an open and transparent mechanism for clearly communicating and documenting System Architecture Working Group (SAWG) synthesis. Thus, the ***scope*** of this report focuses on synthesizing the information provided by workshop #2 participants and documenting a path forward for the SAWG by contextualizing these inputs in the larger study.

2. Reference Documents and Materials

Study Website with links to workshop agendas, presentations, and related documents: https://earthdata.nasa.gov/esds/open-science/oss-for-eso-workshops

Transform to Open Science Github: https://github.com/nasa/Transform-to-Open-Science

Workshop 1 Report: http://hdl.handle.net/2014/53042

3. Executive Summary

The purpose of the ESO Mission Data Processing Study is to identify the architectures that best:

1. Meet the ESO mission science data processing objectives
2. Enable data system efficiencies
3. Support Earth system science and applications, and
4. Promote open science principles to expand participation in mission science beyond the funded science teams

Sponsored by Kevin Murphy, Chief Science Data Officer, NASA Science Mission Directorate (SMD), and Program Manager, Earth Science Division (ESD), it's focused on the four ESO missions: NISAR/SDC, SBG, MC, and AOS. The study team consists of the Steering committee whose primary role is management of the study and the System Architecture Working Group (SAWG) responsible for conducting the Mission Data Processing System (MDPS) architecture trade study. The study is conducted via a series of three workshops. Workshop #1 focused on gathering needs and considerations and was completed in Oct 2021. The goal of workshop #2 was to understand the state of the practice and state-of-the-art in Big Data processing systems. This document is a report of the findings of Workshop #2.

The Workshop was conducted over 4 days with 7 sessions focusing on 1) Science collaboration, 2) NASA Earth missions, 3) NASA Earth Pathfinder missions, 4) Non-NASA Earth missions, 5) Non-Earth missions, 6) System interfaces, and 7) Non-NASA MDPS, and multiple breakout and group discussions. Additionally, a survey was conducted to gather deeper context and information from the presenters and the wider community. A summary of the workshop and the path forward follows.

The workshop was kicked off by the NASA Open Science Initiative, implemented through the TOPS project, which identified as a principal need, the implementation of capabilities that enable and foster Open Science. A major need is a Big data computing platform, for analysis, in an integrated approach. While many computing platforms exist, they are focused on a specific domain and limited by the specific datasets available in the platform.

While the data processing systems that were examined, including the NASA/non-NASA and mission/non-mission, contained common functionalities, they were built upon a wide variety of implementations. The main common components, at the highest level, were the data component and the processing component, which were the main drivers for the architectural decisions and cost implications. Ranging from GBs to PBs for storage needs, and a few to 1000's of processing nodes, the various MDPS's were deployed either solely in an on-prem facility, wholly in a cloud platform (the most popular by far being AWS), or hybrid (on-prem & cloud). Some took advantage of NASA HECC for overflow and reprocessing. Heritage of both software and hardware were prominent across missions implemented within an organization and seen as a leading factor in making subsequent missions cheaper/more efficient. Three main MDPS architectures were identified, 1) Single Instance: one system for one mission, 2) Multi-mission

System: one instance to process multiple missions, 3) Co-located MDPS & DAAC: one system for one mission but sharing functions with the DAAC. A common theme among the non-NASA systems e.g., NOAA, JAXA, ISRO, ASI, DLR, was that they were multi-mission (one instantiation to support multiple missions) and generally built around a data lake, where the NASA missions were generally single instantiations per mission. ESDIS's migration of all the DAAC's data to AWS is a foundational step in constructing a data lake, around which efficient processing and access services are being developed. It was promising to see the trend for non-mission MDPS's focused on big data, analysis platforms, in support of the science community, e.g., cyverse, Pangeo, OpenSARab, OpenNEX, & NASA EIS.

The breakout sessions and open discussion highlighted some open issues. While SPD-41 mandates the policy, the challenges between Open Science and cybersecurity remain unaddressed, especially at the organizational level. Additionally, better processes for community contributions, interoperability, quality assurance, data and metadata standards, and the advancement of capabilities such as ARDs remain a challenge. When examining multi-mission systems, the increased efficiency and support for system science must be weighed against cost management (particularly if it's opened for public use) and interdependency complications. Movement to the commercial cloud is prominent, yet open debate on the benefits and limitations of on-prem versus cloud remain, especially considering cost and capabilities.

With the criteria defined from Workshop #1, and the state of the practice and state of the art in MDPSs survey completed in Workshop #2, the SAWG is ready to embark on the core of the study. Starting with the three common architectures, with additional architectures to be defined from the starting common functionalities of an MDPS (Block Definition Diagrams), the approach will follow traditional NASA System Engineering Handbook processes to conduct the study and present the results in Workshop #3, planned for August 2022.

4. Highlights

- There are a wide variety of implementations of MDPS because of the differences in the environments, data types, and community needs; yet there are common functionalities that an MDPS has.
- Many of the systems acknowledged a movement to commercial cloud, and there is still a very open debate on the benefits and limitations of on-prem versus cloud and what an optimal hybrid option would be that makes the most of both cases (the two are not mutually exclusive).
- While open science is a key objective for use of the data, there are benefits to having some components of an MDPS that have closed access (e.g., project artifacts such as operating environment, software, code, services, data, etc.), while still providing open transparency in the methods for reproducibility, system use metrics, and contribution credit.
- Most MDPS systems are trying to open their system for broader use and some of the common challenges include cybersecurity, data movement, and access.
- MDPS were expanding to accommodate data users who want to produce products as much as they want to use the data for science analysis.
- MDPS are leveraging software and services built by ESDIS and other open-source capabilities (e.g., Jupyter, Docker, Kubernetes, etc.).
- Opening up data and new systems creates new challenges for when a product transitions from an MDPS to an archive (e.g., quality assurance and control) and how an MDPS facilitates reproducibility and provenance.
- Sustainable cost models and cost accounting for a more open science MDPS is not clear.

5. ESO Study Overview

The motivation for a data system architecture study stems from recognition that access to algorithms, workflows, computing, and analytics has been a major barrier to participating in NASA science. Opening the access provides greater opportunities for more people to participate in NASA science.

The purpose of this study is to assess methods to enable data system efficiencies to support the next decade of NASA Earth System Observatory (ESO; Margetta, 2021) missions that support Earth system science and promote open science principles to expand participation in mission science beyond funded science teams. The focus is not specifically on data archiving, but on how the broader community can participate in mission data system processing (see glossary).

The objectives of the data system architecture study are to identify and assess potential data system architectures that can:

1. Meet the ESO mission science data processing objectives
2. Enable data system efficiencies
3. Support Earth system science and applications
4. Promote open science principles to expand participation in mission science beyond the funded science teams

The principles of this study are to practice open, team science by conducting meetings in the open, thoroughly recording the conversations during workshops, and by making workshop artifacts citable with DOIs and accessible through the Study website and Github. In addition, participants in one-on-one meetings (between the study team and other entities) should document and make notes from the meetings accessible, ensuring community participation, provide mechanisms for continuous feedback, and actively seek feedback from historically excluded communities.

This study is sponsored by Kevin Murphy, Chief Science Data Officer, Science Mission Directorate, and Program Manager, Earth Science Data Systems Program. The study consists of core staff who help gather inputs from a broader community including the Steering Committee and the SAWG. Descriptions of the roles of these core staff are outlined in the glossary. The SAWG is responsible for collecting and evaluating data system architecture drivers including: 1) ESO program goals, constraints, and opportunities; 2) ESO mission objectives and capability needs; 3) the state of the practice in open-science and data processing systems; and 4) community recommendations. The SAWG will perform a trade study that establishes viable architectural options and implementation approaches. To accomplish this, they will establish evaluation criteria (qualitative and quantitative) for use in the analysis of the trade space. The

SAWG will use the trade study to provide candidate architectures and make recommendations. All methods and findings will be clearly documented.

The approach of the study is to solicit stakeholder feedback through open workshops and public Requests for Information (RFI). There are three workshops planned. Workshop #1 focused on understanding the NASA program goals and ESO mission needs with the explicit goal of informing both qualitative and quantitative evaluation criteria of different architectures against the open source science data system objectives. It was held virtually on Oct 19-20, 2021. Workshop #2 focused on understanding the state-of-the-art in mission data processing systems and open science, as well as seeking community input on data system architectures. It was held virtually on March 1-4, 2022. After Workshop #2, the SAWG will conduct a system architecture trade study evaluating different architectures against the evaluation criteria. Over the 4-month period during this study, the SAWG communicated with stakeholders from Workshop #1 to inform assessment of architectures for meeting different criteria. Workshop #3 will present candidate architectures and make a recommendation with an assessment against the criteria. It is planned for August 2022 in a virtual format. NASA Headquarters will then decide a path forward.

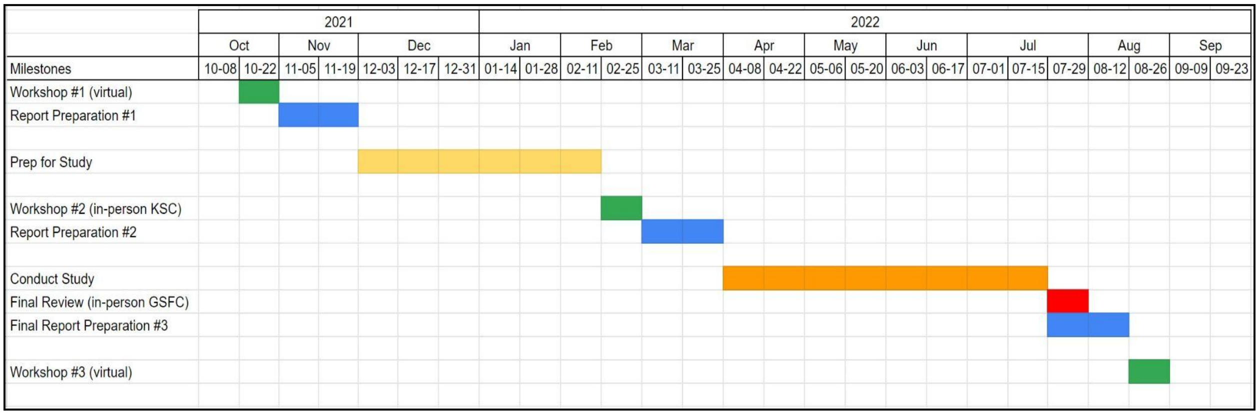


Figure 1. A gantt chart of the project timeline.

6. Workshop #2 Format and Overview

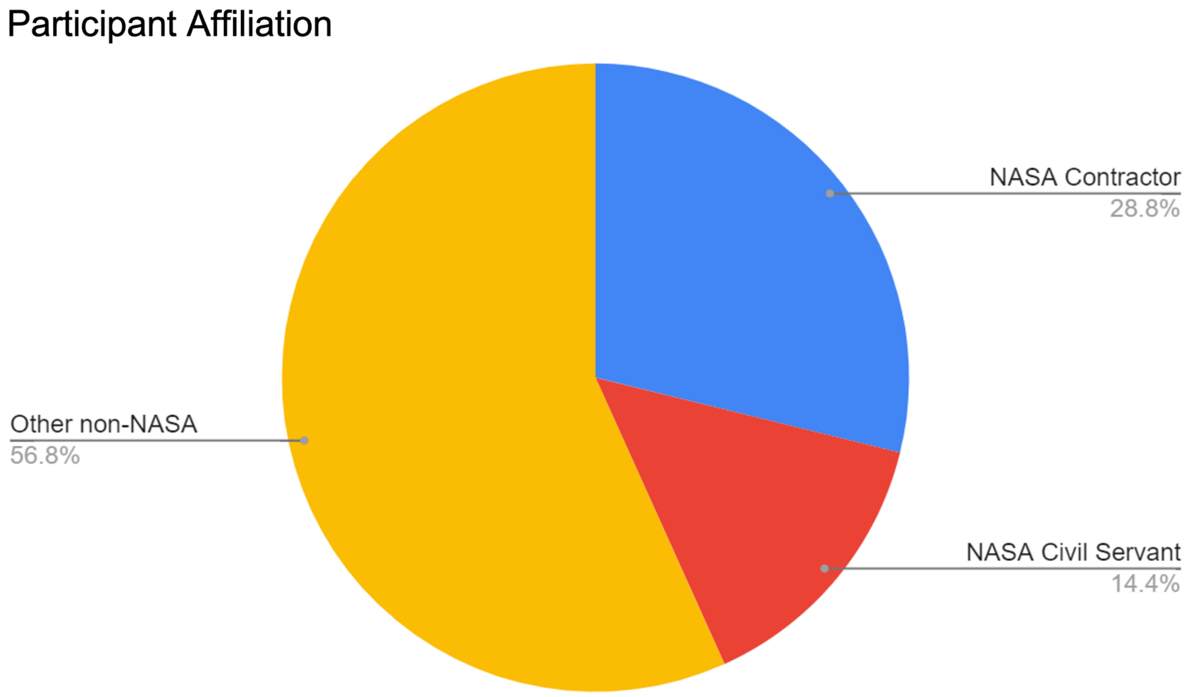
This workshop focused on understanding the state of the practice and state-of-the-art in Big Data processing systems and was open to public participation through registration. A Request For Information (RFI; See Appendix) sought input organizations outside of NASA with relevant expertise in Big Data processing and open science to participate in select sessions, and the study and workshop discussions.

The 4 Day workshop (March 1-4, 2022) agenda included 7 specific sessions, with talks grouped to guide discussion and facilitate information gathering for the SAWG. The individual sessions were: 1) Science collaboration approaches, 2) NASA Earth System mission processing, 3) NASA Earth System Science Pathfinder mission processing, 4) Non-NASA Earth science mission processing, 5) Non-Earth science mission processing, 6) System interfaces and standards, and 7) Non-NASA MDPS. Each session concluded with a “Fishbowl Discussion” as a time for the SAWG members to dialogue with and direct questions to the speakers.

Days 1-2 ended with focused breakout discussions in virtual breakout rooms and covered six topics: 1) System development approaches and challenges, 2) System operations approaches and challenges, 3) Open-sourced science approaches and challenges, 4) Data analysis needs for a Mission Data Processing System (MDPS), 5) Open source software approaches and challenges, and 6) MDPS Architectures now and the future). These sessions were led by a SAWG member discussion facilitator and separate note taker. Those in the breakout synthesized their discussion in a summary of key take-away points to represent the discussions.

On Day 3 used mentimeter to facilitate a more informal big group discussion and identify topics that had not yet been discussed. This resulted in organic conversations about defining an MDPS and the utility of on-premise scalable compute versus commercial cloud options.

Workshop participants included 437 people registered for the 4 day workshop representing government, academia, private industry both domestically and internationally. The attendance rate varied by day.



To ensure everyone (speakers and workshop attendees) had the opportunity to participate, a survey was provided with questions about MDPS implementations, deployments, operations, interfaces with NASA and analysis platforms as well as ways to incorporate community contributions for an open-source science data processing system. This survey enabled all participants to answer the same questions in a systematic way to understand commonalities and differences. Questions collected both multi-choice categorical data as well as free-form written responses.

7. Workshop #2 Findings

The findings from this workshop include both those from a 15-minute, optional survey sent around to all speakers and participants (Section 7.1), and a synthesis of what was presented by speakers, who were given templates with similarly inspired questions to the survey, but an opportunity to show diagrams and discuss nuances specific to each MDPS (Section 7.2). These speakers fell into seven categories: 1) Open Science and Collaboration (Section 7.2.1), 2) NASA Earth System Missions (Section 7.2.2), 3) NASA Earth Pathfinder Missions (Section 7.2.3), 4) Non-NASA Missions (Section 7.2.4), 5) Non-Earth Missions (Section 7.2.5), 6) NASA System Interfaces (Section 7.2.6), and 7) Non-NASA MDPS (Section 7.2.7).

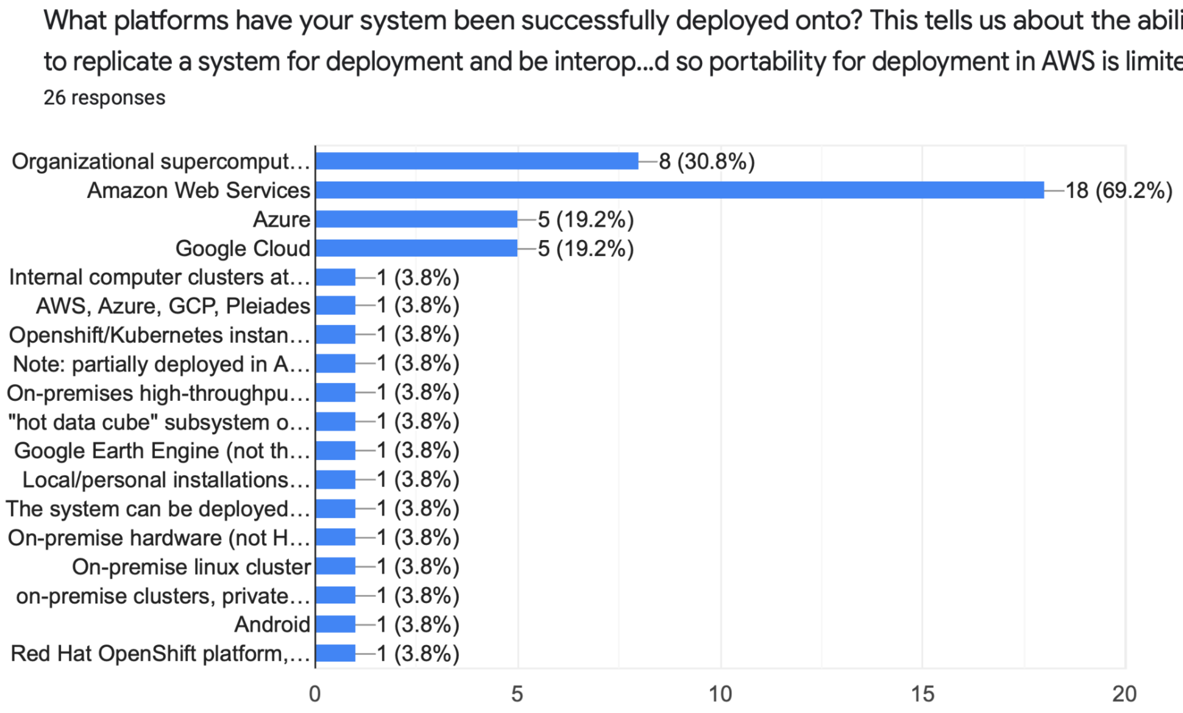
7.1 Survey and Results

We received 30 responses to the survey. Most of the responses (53.6%) were from government personnel, followed by non-profit (25%), academia (17.9%), and Industry (3.5%). Most of the responses (75%) were from integrated processing systems and about 11% were from component technologies or services. The primary customers of these responders are public, followed by in-house requests and sponsored principal investigators. Responders split almost evenly between operational and research/development.

Most of the components/services are of high TRL: more than 50% with a TRL greater than 7, and 32% with TRLs of 5 and 6. About 65% of the responders, each representing a different MDPS, support interoperability and indicated that their architectures are using open-source software. Over 70% of the survey respondents indicated that their systems already interface with NASA, with most leveraging legacy processing systems based upon open-source software.

About 40% of the systems included in this survey were described as having unique requirements/constraints (e.g. time sensitive data, instrument support/responsibilities, resource, etc.). 32% of the systems do not have these requirements, and the rest indicated that they partially have these requirements.

Funding for these systems is mainly supported by projects/programs, though some respondents indicated they did not know their cost models between shared and single project funding schemes leveraging different scalable compute platforms from high-performance computing (HPC) and different commercial cloud vendors. Close to 60% of the architectures rely on service components that are shared across projects, while the other 40% are solely owned by a single project. The most common usage of the system is by a single project for a single purpose (23%), followed by multiple organizations with a variety of purposes (19%), and by one organization with a specific purpose (15%). Most of the systems (78%) indicate that it is possible for multiple projects to use a shared service, while the rest indicate their systems can partially support multiple projects. Almost 70% of the systems have been successfully deployed in AWS and 30% onto organizational supercomputer clusters. Some of these systems have also been deployed in multiple cloud services, including Azure (~20%) and Google cloud (~20%), and other platforms. When operating and maintaining different systems, the most common hidden costs are compute, data storage, workforce, and software development.

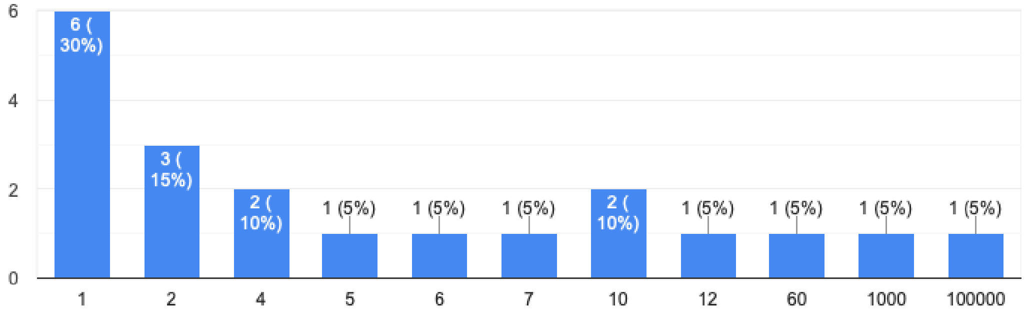


What platforms have your system been successfully deployed onto? This tells us about the ability to replicate a system for deployment and be interop...d so portability for deployment in AWS is limited.

26 responses

Nearly half of these systems require multiple containers to execute the processing workflow; approximately 8% of the systems indicated that a single container is sufficient, and about 38% of systems can do both. A full 63% of the systems incorporate more than a single cloud provider and/or use on-premise systems.

All but one responder indicated that software and product upgrades are version controlled. Only about 30% of the systems can easily accommodate additions/substitutions. There is large spread in terms of instantiations of the systems:



Most of the current systems do not have a formal venue to support community demand for variations of products. A few use github and open community forums to facilitate a dialogue with the user community, while others rely on international organizations, although some systems indicate there is no need for this communication. Some systems provide web-based tools for users to generate on-demand products, while many of the systems do not support on-demand product generation.

More than half (54%) of the systems provide interfaces with public-facing analysis platforms. Most of the systems (96%) indicated that they can be modified to incorporate interfacing with public-facing analysis platforms. For those systems that interface with public-facing analysis platforms, 82% of them also indicate that their systems could allow for community contributions to improve, augment, and/or modify their systems. Most of these systems accept community contributions via open forums such as GIT repository, a few via email or online requests. Additionally, about 57% of those systems can incorporate and/or run some community contribution developed on the public-facing platform. The community can contribute via GitHub or Docker container, some projects require that contributions adhere to their processing system specification, and some require that community contributions adhere to open and interoperable standards.

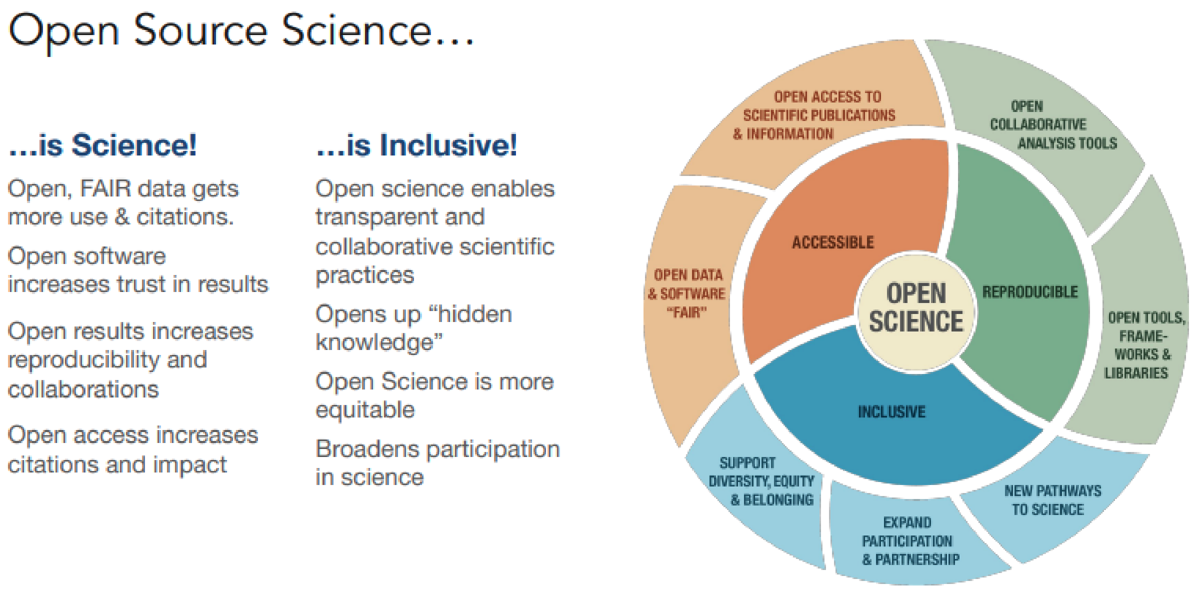
7.2 Summary from Speakers

7.2.1 Open Science and Collaboration

Expanding participation, improving reproducibility, and accelerating scientific discovery for societal benefit

NASA's strategy for data management and computing for groundbreaking science (NASA, 2019) sets forth a new vision to enable transformational open science through the continuous evolution of science data and computing systems. The core mission of the strategy is to lead an innovative and sustainable program that supports NASA's unique science missions with academic, international, and commercial partners and to continually evolve systems to ensure they are usable and support the latest analysis techniques while protecting scientific integrity.

With the strong desire to promote accessibility of Scientific Information within the NASA-funded community and the general public, a principal need is the implementation of capabilities that enable and foster Open Science. In general, Open Science can be loosely defined as a collaborative culture enabled by technology that empowers the open sharing of data, information, and knowledge within the scientific community and the wider public. By building on concepts from Open-Source Software, which greatly expanded participation in developing code, we can extend these concepts to the scientific process, which may accelerate the discovery of scientific discovery by openly conducting science from project initiation through implementation. Merging the two concepts of Open Science and Open-Source Software, the new term Open-Source Science (OSS) is defined (see Glossary).



Open Source Science…

…is Science!

Open, FAIR data gets more use & citations.

Open software increases trust in results

Open results increases reproducibility and collaborations

Open access increases citations and impact

…is Inclusive!

Open science enables transparent and collaborative scientific practices

Opens up "hidden knowledge"

Open Science is more equitable

Broadens participation in science

OPEN ACCESS TO SCIENTIFIC PUBLICATIONS & INFORMATION
OPEN COLLABORATIVE ANALYSIS TOOLS
OPEN DATA & SOFTWARE "FAIR"
ACCESSIBLE
OPEN SCIENCE
REPRODUCIBLE
OPEN TOOLS, FRAME-WORKS & LIBRARIES
INCLUSIVE
SUPPORT DIVERSITY, EQUITY & BELONGING
EXPAND PARTICIPATION & PARTNERSHIP
NEW PATHWAYS TO SCIENCE

A consistent Open-Source Science Policy that is clearly communicated to the community is important, therefore the Scientific Information Policy SPD-41 (NASA, 2021) has been created and is now widely distributed. The SPD-41 policy highlights four core values 1) preservation of scientific information, 2) maximizing openness of scientific information while 3) minimizing the burden on the community in complying with the policy, and 4) growing the community that can access NASA’s scientific information. “Scientific Information" is broadly defined to include publications, data, and software following these guidelines: an open-access version of the as-accepted manuscripts must be accessible via a NASA designated repository; data shall be made publicly available without fee or restriction of use and no period of exclusive access; and software should be released as open-source software with a permissive license.

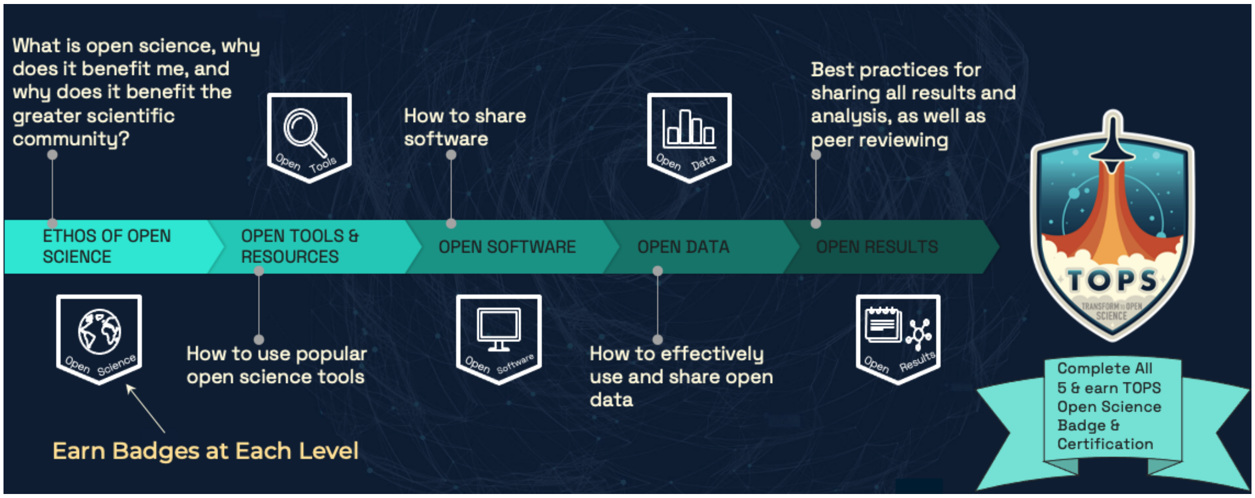
It should be noted that all future awards will be in compliance with SPD-41 and all proposing entities should plan accordingly and budget for any additional work imposed on the proposing team. A proposed addition to SPD-41, SPD-41a, will include Findability, Accessibility, Interoperability, and Reusability (FAIR) principles (Wilkinson et al., 2016).

Science collaboration approaches

NASA is taking an approach that facilitates: 1) open transparent science – i.e., a scientific process that makes data, tools, software, documentation, and publications findable, accessible, interoperable, and reproducible (FAIR; Wilkinson et al., 2016); and 2) an open and inclusive process of participation and collaboration with diverse people and organizations.

NASA will lead a path to open science with the Transform to Open Science (TOPS) initiative. This initiative is a $40 million USD, 5-year (pending appropriations) program within the Science Mission Directorate with the stated objectives of increasing understanding and adoption of open science, accelerating major scientific discoveries, and broadening participation by historically underrepresented communities. Several incentives are being considered with TOPS such as the Open Science Prizes and Awards program, which may reward those that demonstrate significant leadership and progress toward open science and showcase the benefits of open science. Other existing award programs are being reviewed and updated to include open science activities as review criteria.

NASA 's proposed plan is to use the year 2023 for the Year Of Open Science (YOOS), which will announce sweeping changes across funding decisions, awards, promotions, evaluations, and the recognition of teams as well as individuals. YOOS will require open FAIR data (Wilkinson et al., 2016), open software, open access publications, and a pathway to earn open science badges that funding awards and promotion decisions can consider as part of review criteria. A badge is shown and defined in the following figure:



What is open science, why does it benefit me, and why does it benefit the greater scientific community?

How to share software

Best practices for sharing all results and analysis, as well as peer reviewing

ETHOS OF OPEN SCIENCE

OPEN TOOLS & RESOURCES

OPEN SOFTWARE

OPEN DATA

OPEN RESULTS

How to use popular open science tools

How to effectively use and share open data

Earn Badges at Each Level

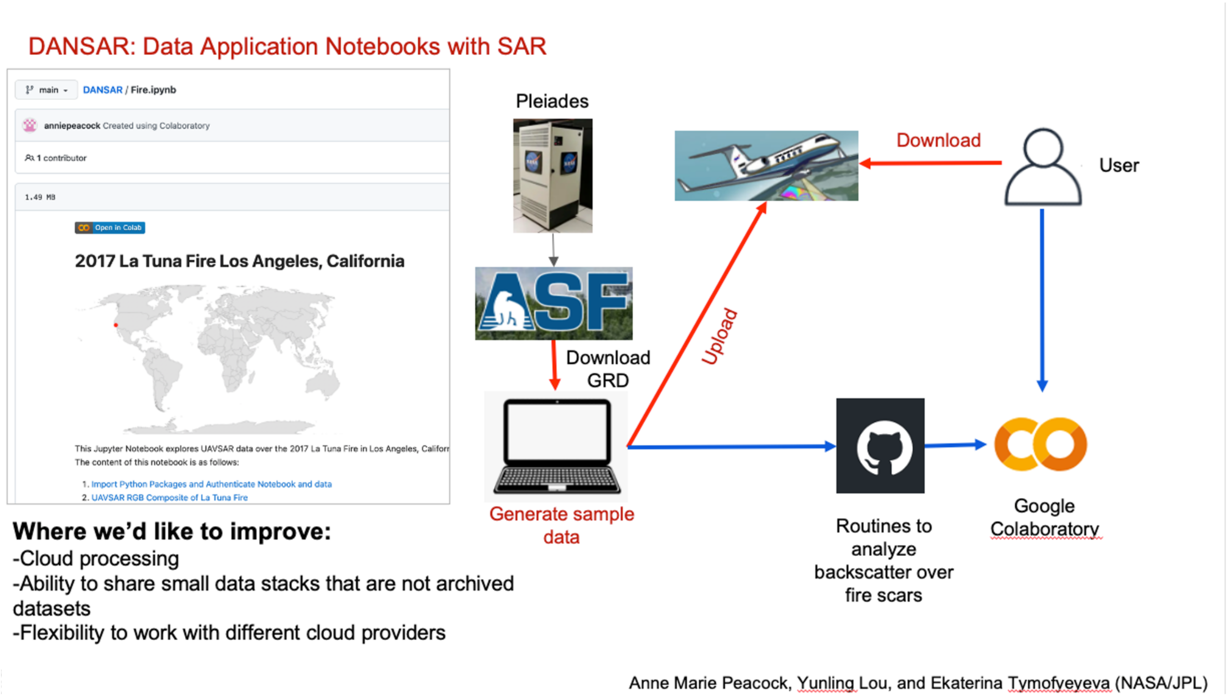
Complete All 5 & earn TOPS Open Science Badge & Certification

Volunteers are needed to implement this cultural shift, and community engagement from the broad spectrum across the scientific community is required.

SAR Data Exploration - Challenges and Opportunities in the Environmental Sciences

SAR data has utility for a diverse set of customers and stakeholders, from scientists and decision-makers within US federal, state, and international agencies, to academic researchers, all with varying skill sets. While customers need synoptic maps of land cover, carbon content, inundation status, and other terrestrial biophysical parameters, they also need to integrate SAR with other data types, such as optical and in-situ observations.

What's needed is a mechanism or platform for the user community to explore and analyze data with this integrated approach. Many hurdles exist for the community to work with the large data volumes from SAR, produce usable products, and conduct analyses. Specifically, one such barrier is the lack of adequate infrastructure including high-speed internet access, appropriately resourced systems with Linux operating systems (particularly outside of the USA), and robust, scalable Geographic Information Systems (GIS) clients. One approach is a move towards a notebook-first environment with Jupyter notebooks, hosted on cloud platforms (e.g., Google Cloud).



DANSAR: Data Application Notebooks with SAR

main DANSAR / Fire.ipynb

anniepeacock Created using Colaboratory

1 contributor

1.49 MB

Open in Colab

2017 La Tuna Fire Los Angeles, California

This Jupyter Notebook explores UAVSAR data over the 2017 La Tuna Fire in Los Angeles, Califor
The content of this notebook is as follows:

1. Import Python Packages and Authenticate Notebook and data
2. UAVSAR RGB Composite of La Tuna Fire

Pleiades

Download GRD

Generate sample data

Upload

Download

User

Routines to analyze backscatter over fire scars

Google Colaboratory

Where we'd like to improve:
-Cloud processing
-Ability to share small data stacks that are not archived datasets
-Flexibility to work with different cloud providers

Anne Marie Peacock, Yunling Lou, and Ekaterina Tymofyeyeva (NASA/JPL)

While the mechanics of collaborating on non-standard data or research products is possible, and the team is demonstrating this today, doing so at scale is difficult from a cost perspective, i.e. cloud storage and data egress costs. Better data governance may be needed to move forward on facilitating collaboration with data at scale, and the team looks to NASA to address cost implications, governance, etc. Additionally, NASA's current cybersecurity posture does not map well to the current procedures and/or protocols around the concept of sharing early and sharing often, a goal of NASA's Open Source Science Initiatives.

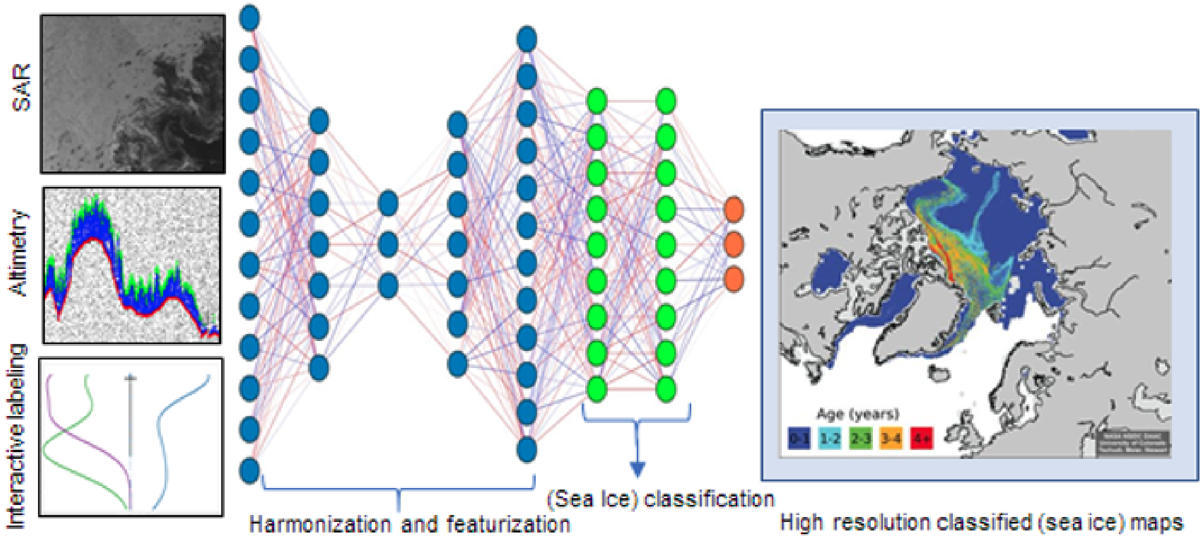
Imagining a new NASA computing platform with AI + big data-supported analysis

There are several commercially available systems, such as Google Earth Engine (GEE; https://earthengine.google.com/), that provide analytics, machine learning, and general data processing capabilities that may help NASA promote access to Scientific Information. It was reported that GEE has clear documentation, online GUIs for ease of use, particularly for non-experts, good, easy integration with Google Colab (https://colab.research.google.com/) and a wide range of cached datasets which may accelerate general data analytics, machine learning, and data munging tasks for users of NASA's Scientific Information. There were several pain points for using the GEE platform; specifically a lack of NASA's dataset cached in GEE and the difficulty to import into and cache new datasets on the platform as of today. There are some concerns about costs to use the platform for users with limited budgets, which could damper NASA's goal of inclusiveness unless this is acknowledged and steps are taken to ensure some form of long-term, low-cost access to the platform. Linkages to code repository platforms, such as GitHub, may help with the mechanics of disseminating and/or using open source software, a key interest in the NASA OSS initiatives.

On reimagining data access and processing, it may be desirable to provide a unified web-based compute platform where all data is collocated and a number of elements are included in the platform. At a high level, the platform should support multiple user skill levels, enable a mechanism for real-time user discussion, provide a mix of hardware types such as GPU, promote open source software, collaborations and provide the ability to export final products (data sharing). This reimagines the current NASA Distributed Active Archive Centers (DAAC) architecture, which is challenging for some users, as the data of interest for a particular study may be stored at multiple DAACs, making it harder to find, acquire, and munge files that may not be "analysis-ready", an already tedious pain point.

Big Data Community Algorithms: Deep Learning for Mapping

Geoscientists often require working with a diverse set of data when developing advanced machine learning methods to harmonize heterogeneous Earth observations for classification and mapping, such as sea ice. Providing user-friendly data management and machine learning-based pipelines which can leverage data in the cloud with ease is highly desirable.



The geoscientist community can benefit from accessible, shared, low-cost, vendor-managed processing capabilities on a variety of hardware architectures, such as GPU-enhanced nodes and high memory nodes. It was suggested that domain scientists working with, and presumably attempting to manage, traditional computing clusters can be a significant barrier or burden.

One of the challenges for the machine learning community is that there is a dearth of training data and a lack of pre-trained models for this specific community. There is a strong interest in building and sharing standard pre-trained models, which are tedious to create due to the lack of pre-canned multi-sensor, multi-mission, spatiotemporally harmonized data products. Mechanisms to facilitate incorporating domain expertise into the scientific process could benefit various activities for researchers with varying levels of experience.

Cloud Computing Platforms for Processing Geospatial Big Data: Current Status and Challenges

There are a variety of computing platforms for working with big geospatial data. Currently, platforms of interest are Google Earth Engine (https://earthengine.google.com/), Microsoft Planetary Computer (https://planetarycomputer.microsoft.com), Pangeo Cloud (https://pangeo.io/cloud.html), Sentinel Hub (https://www.sentinel-hub.com), OpenEO (https://openeo.org), Open Data Cube (https://www.opendatacube.org), and the Multi-Mission Algorithm and Analysis Platform (MAAP; https://scimaap.net). Each platform has unique features, available programming languages, application programming interfaces (APIs), datasets directly provided (cached) or accessible, data cataloging, and tooling. Platforms may focus on specific communities and have assumptions about the level of both domain and technical expertise.

Platforms may support several platform backends (a platform of platforms), OpenEO is one such example. OpenEO provides a unified way to access several publicly accessible backends such as Google Earth Engine and Sentinel Hub as well as other private backends. There may be a potential benefit of providing a mechanism or unified way to move a specific task to the locality of the data and/or the specific platform which provides the best tool or API for the task. A factor that might limit the ability to easily move tasks from platform to platform is the lack of data standards and/or protocols.

General needs and capabilities for any platform are to facilitate reproducibility, replicability, interoperability, scalability, and financial sustainability, while being extensible, easy-to-use, and fostering the use of open-source software.

Project Jupyter - Lessons and Principles from a Community-Driven Open Source Project

Jupyter (https://jupyter.org/) is an open community dedicated to modular, platform-agnostic tools for interactive computing. The Jupyter community has a formal governance structure with support from several institutional partners and sponsors. Jupyter is more than just software, additional areas of focus are services, content, standards and protocols, and the community.

JupyterLab is a modular architecture with a focus on the needs of the data science community. The JupyterLab module toolkit addresses each of the elements within the life-cycle of research. Ideas may be noted with these elements: individual exploration, collaborative development, large-scale production runs (in High Performance Computing (HPC) and Cloud), publishing and communicating results, and education.

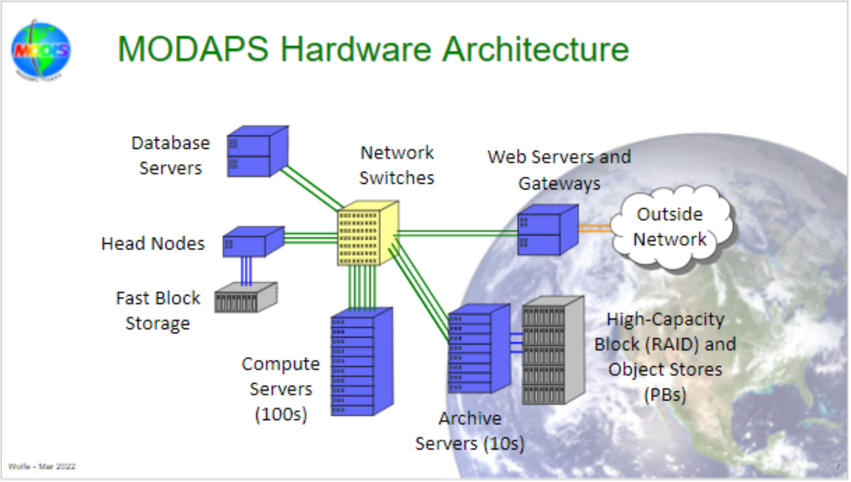
With improved researcher productivity through ease of use, accessibility, and deployability of the tools developed and maintained within the Jupyter ecosystem, impacts on both research and education have been significant to date. The breadth and volume of impactful science and education are expected to increase as coursework around data science using tooling, such as JupyterLab, grows.

The platform-agnostic nature of the module tooling in the Jupyter ecosystem may guard against vendor lock-in and this cloud-neutral approach may provide resilience, reduce the risk of a single point of failure, or unintended exclusion due to a change in cloud pricing models as an example. Furthermore, combining datasets that are hosted across cloud service providers is highly desirable, as requirements for unique hardware (TPU, GPU, etc.) for specific problems is becoming more commonplace. Allowing the community to leverage each cloud vendor's strengths and weaknesses is advantageous and should be considered.

7.2.2 NASA Earth System Missions

Terra MODIS Instrument

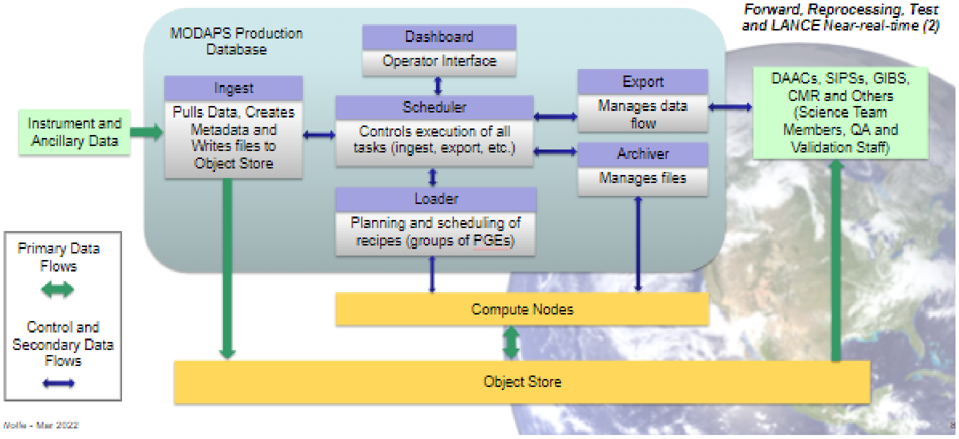
There are two Moderate Resolution Imaging Spectroradiometer (MODIS) instruments flown on Aqua, launched in 2002 and the second on Terra, launched in 2019. The MODIS Science Investigator-led Data System (SIPS) is called the MODIS Adaptive Processing System (MODAPS) and is collocated with LAADS DAAC. Each MODIS instrument generates ~180 TB of L0, L1, and L2+ land/atmosphere data yearly. MODAPS also produces the Visible Infrared Imaging Radiometer Suite (VIIRS) L1/land products. MODAPS supports near-real-time products, standard mission data processing, and reprocessing with early integration of calibration/validation data and user feedback to improve validation efforts. The MODIS hardware architecture is depicted in the following figure:



The MODAPS processing software architecture is illustrated below:

MODAPS Processing Software Architecture

Multiple MODAPS instances: Forward, Reprocessing, Test and LANCE Near-real-time (2)



MODAPS Production Database

Dashboard
Operator Interface

Instrument and Ancillary Data

Ingest
Pulls Data, Creates Metadata and Writes files to Object Store

Scheduler
Controls execution of all tasks (ingest, export, etc.)

Export
Manages data flow

Archiver
Manages files

DAACs, SIPSs, GIBS, CMR and Others (Science Team Members, QA and Validation Staff)

Loader
Planning and scheduling of recipes (groups of PGEs)

Compute Nodes

Object Store

Primary Data Flows

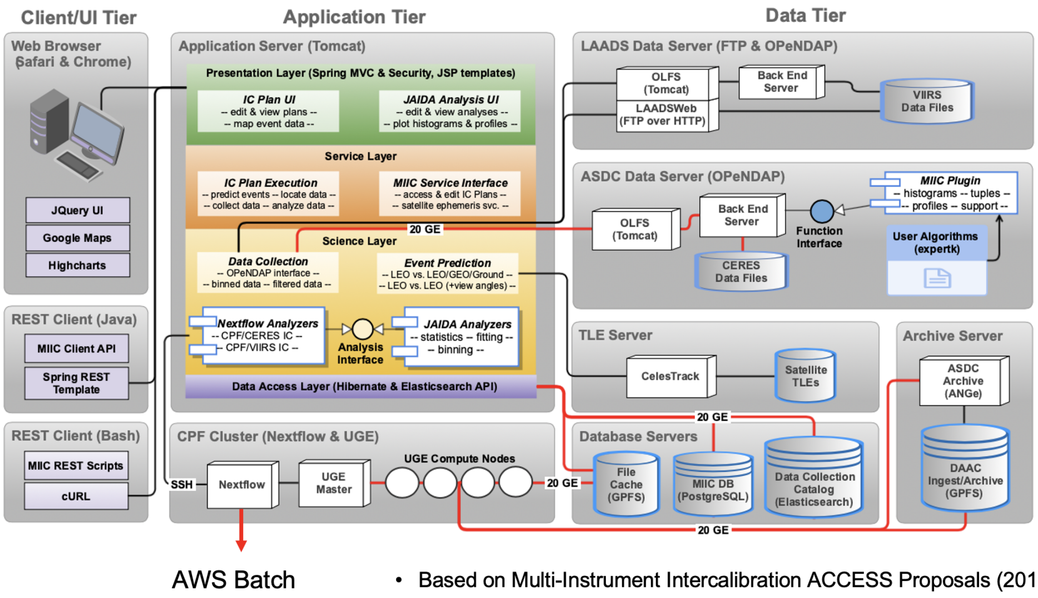
Control and Secondary Data Flows

Wolfe - Mar 2022

CLARREO

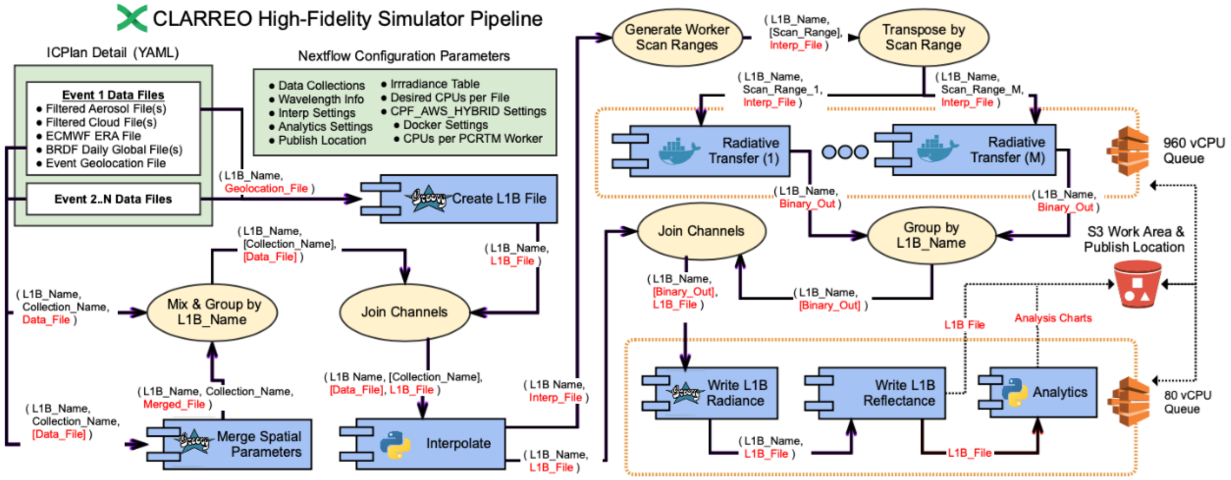
The Climate Absolute Radiance Refractivity Observatory (CLARREO) Pathfinder experiment consists of a reflected solar spectrometer planned for flight on the International Space Station (ISS) in 2023. The instrument has two objectives: (1) demonstrate on-orbit calibration ability to reduce reflectance uncertainty by a factor of 5-10 times compared to the best operational sensors on orbit, and (2) demonstrate ability to transfer calibration to other key satellite sensors by inter-calibrating with CERES and VIIRS. The current hybrid development system architecture is provided below:

Current Hybrid Development System Architecture



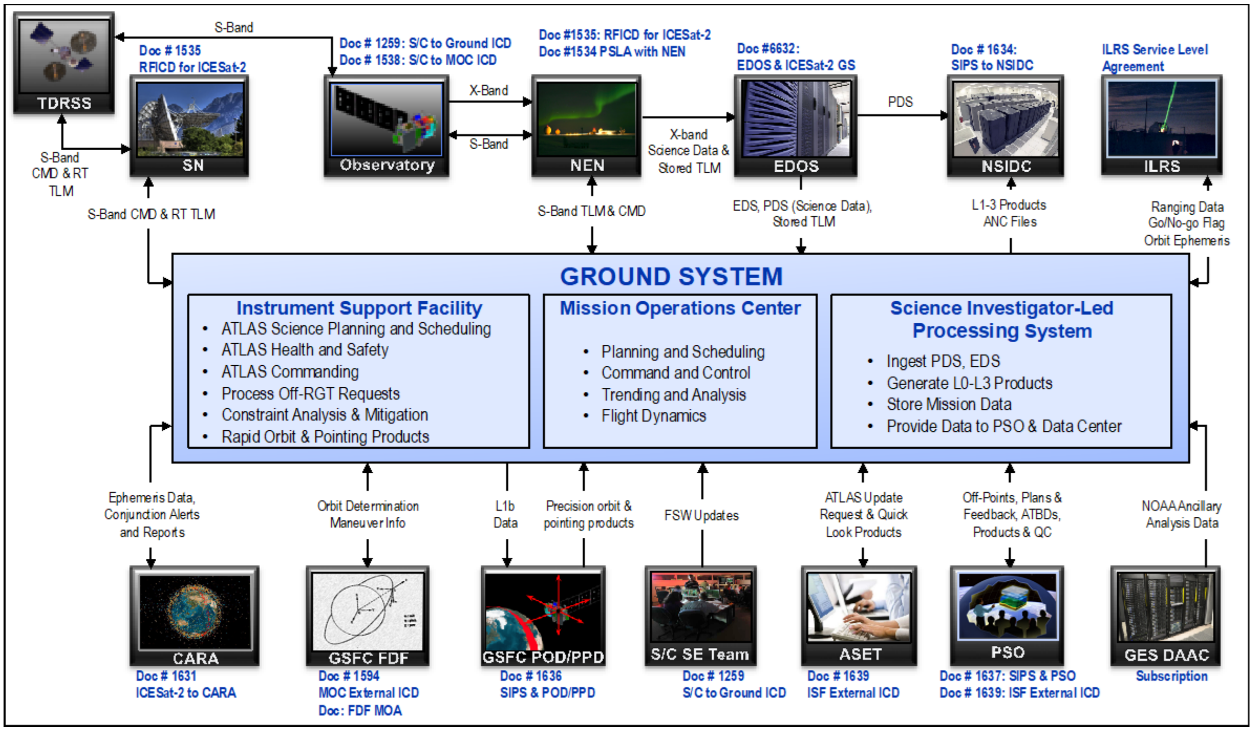
- Based on Multi-Instrument Intercalibration ACCESS Proposals (2011, 2013)

The CLARREO Pathfinder Level 1 high-fidelity simulation using Nextflow pipelines in AWS is provided below:

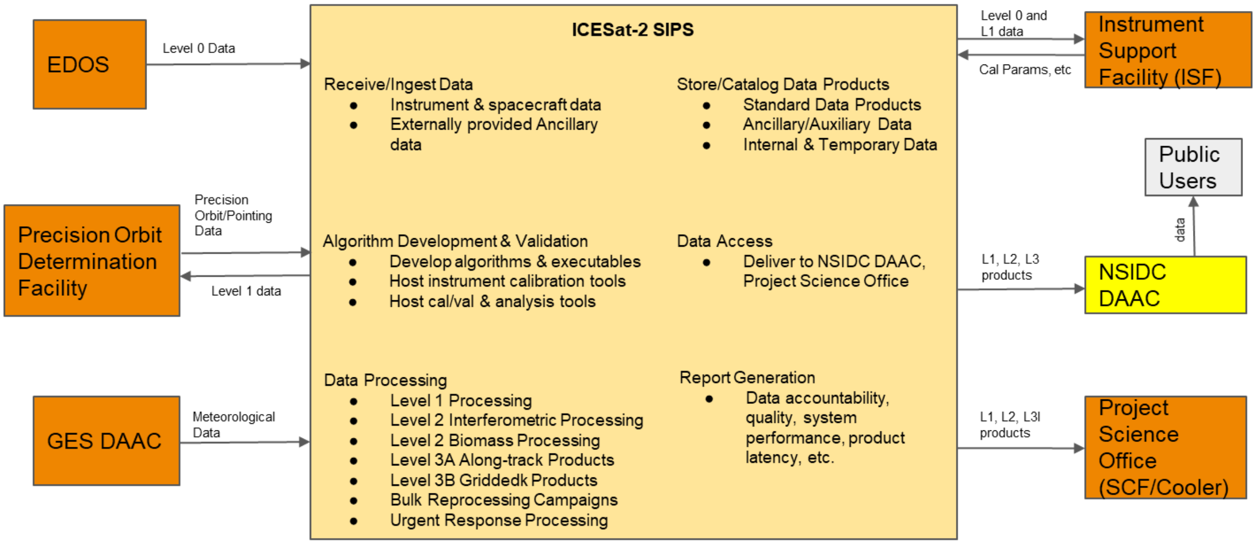

CLARREO High-Fidelity Simulator Pipeline

ICESat-2

The ICESat-2 spacecraft, with a single Advanced Topographic Laser Altimeter System (ATLAS) instrument, was launched in 2018. The instrument generates six laser beams to measure the changing height of ice, clouds, and land elevations. The ICEsat-2 ground system context diagram with interfaces is provided below:



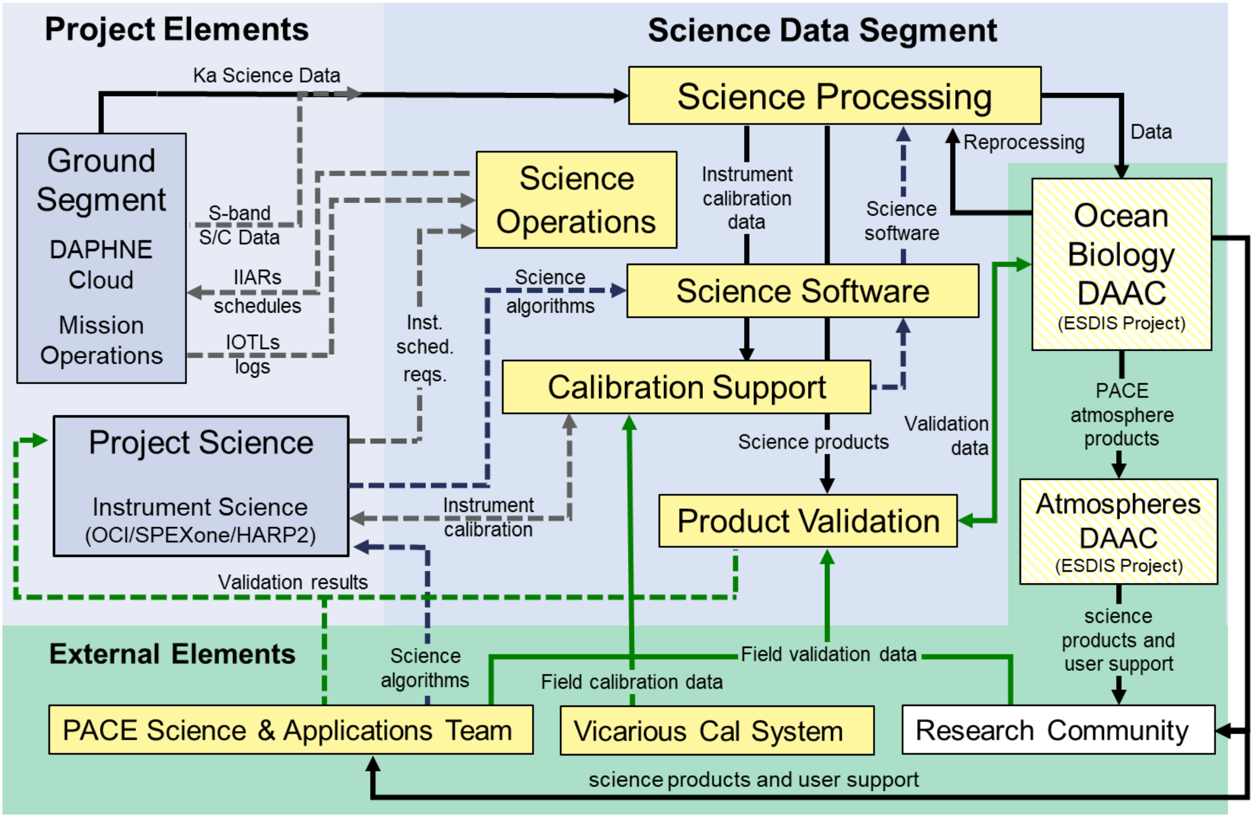
The ICESat-2 data processing system system architecture is provided below:



EDOS
Level 0 Data
ICESat-2 SIPS
Receive/Ingest Data
Instrument & spacecraft data
Externally provided Ancillary data
Store/Catalog Data Products
Standard Data Products
Ancillary/Auxiliary Data
Internal & Temporary Data
Level 0 and
L1 data
Cal Params, etc
Instrument Support Facility (ISF)
Public Users
data
Precision Orbit Determination Facility
Precision
Orbit/Pointing
Data
Level 1 data
Algorithm Development & Validation
Develop algorithms & executables
Host instrument calibration tools
Host cal/val & analysis tools
Data Access
Deliver to NSIDC DAAC, Project Science Office
L1, L2, L3
products
NSIDC DAAC
GES DAAC
Meteorological
Data
Data Processing
Level 1 Processing
Level 2 Interferometric Processing
Level 2 Biomass Processing
Level 3A Along-track Products
Level 3B Griddedk Products
Bulk Reprocessing Campaigns
Urgent Response Processing
Report Generation
Data accountability, quality, system performance, product latency, etc.
L1, L2, L3I
products
Project Science Office (SCF/Cooler)

PACE

The Plankton, Aerosol, ocean Ecosystem (PACE) spacecraft is scheduled to be launched in January 2024. The science products include ocean color, aerosols, cloud optical properties and polarimetry generated by three instruments: Ocean Color Instrument (OCI-hyperspectral radiometer), Hyper Angular Rainbow Polarimeter (HARP-2), and Spectro-polarimeter for planetary exploration (SPEXone-multi-angle polarimeters). The PACE science data segment architecture is depicted below:

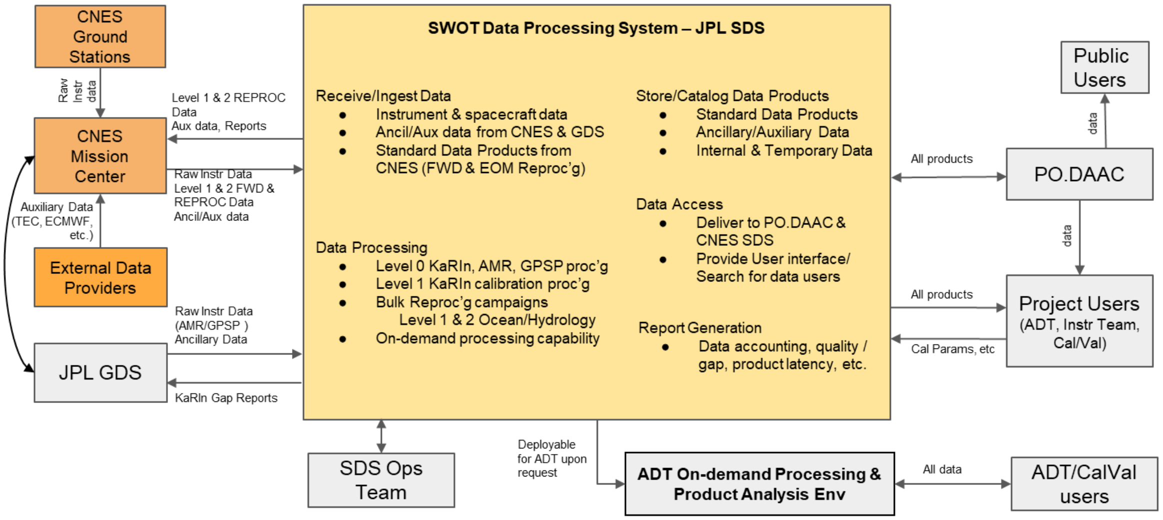


The responsibility for the PACE science data system is directed by the Goddard Space Flight Center (GSFC) Ocean Biology Processing Group (OBPG). All OBPG science data processing codes are open source and made freely available to the public through SeaDAS

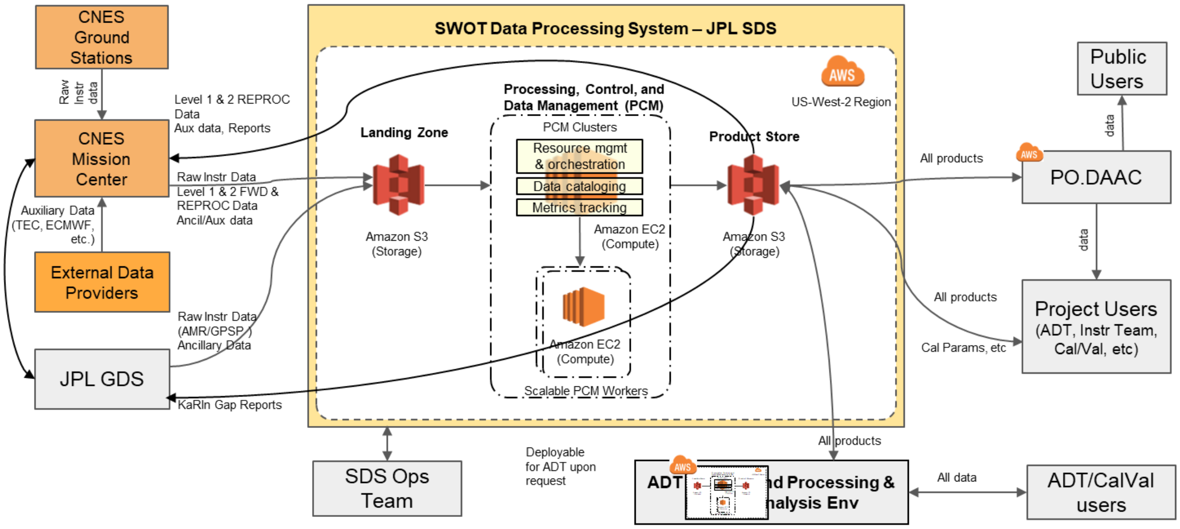
2SWOT

The Surface Water Ocean Topography (SWOT) mission is under joint development by NASA and The French Space Agency - Le site du Centre national d'études spatiales (CNES). SWOT includes contributions from the Canadian Space Agency (CSA) and the United Kingdom Space Agency (UKSA). The mission is scheduled for launch in November 2022 on a Falcon-9 rocket from Vandenberg. Development, testing, and the OPS environments are all deployed in AWS US-West-2 region. The primary instrument is the Ka-band Radar Interferometer (KaRIn).

The SWOT data processing system architecture is depicted below:


CNES Ground Stations
Raw Instr data
CNES Mission Center
Level 1 & 2 REPROC Data
Aux data, Reports
Raw Instr Data
Level 1 & 2 FWD & REPROC Data
Ancil/Aux data
Auxiliary Data (TEC, ECMWF, etc.)
External Data Providers
Raw Instr Data (AMR/GPSP) Ancillary Data
JPL GDS
KaRIn Gap Reports
SWOT Data Processing System – JPL SDS
Receive/Ingest Data
Instrument & spacecraft data
Ancil/Aux data from CNES & GDS
Standard Data Products from CNES (FWD & EOM Reproc'g)
Data Processing
Level 0 KaRIn, AMR, GPSP proc'g
Level 1 KaRIn calibration proc'g
Bulk Reproc'g campaigns
Level 1 & 2 Ocean/Hydrology
On-demand processing capability
Store/Catalog Data Products
Standard Data Products
Ancillary/Auxiliary Data
Internal & Temporary Data
Data Access
Deliver to PO.DAAC & CNES SDS
Provide User interface/ Search for data users
Report Generation
Data accounting, quality / gap, product latency, etc.
Public Users
data
All products
PO.DAAC
data
All products
Project Users (ADT, Instr Team, Cal/Val)
Cal Params, etc
SDS Ops Team
Deployable for ADT upon request
ADT On-demand Processing & Product Analysis Env
All data
ADT/CalVal users

The SWOT component and infrastructure view is provided below:



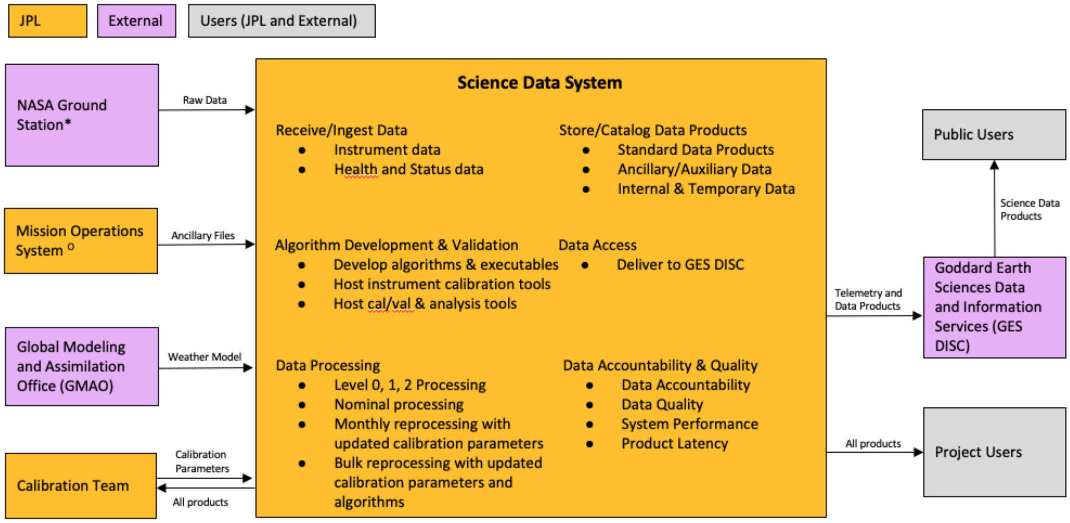
All SWOT products are publicly distributable and algorithm software is archived at the PO.DAAC. Main external factors constrain SWOT Science Data System (SDS) is the need for computation capacity and the collocation of SDS and PO. DAAC to reduce data egress costs and limit.

7.2.3 NASA Earth Pathfinder Missions

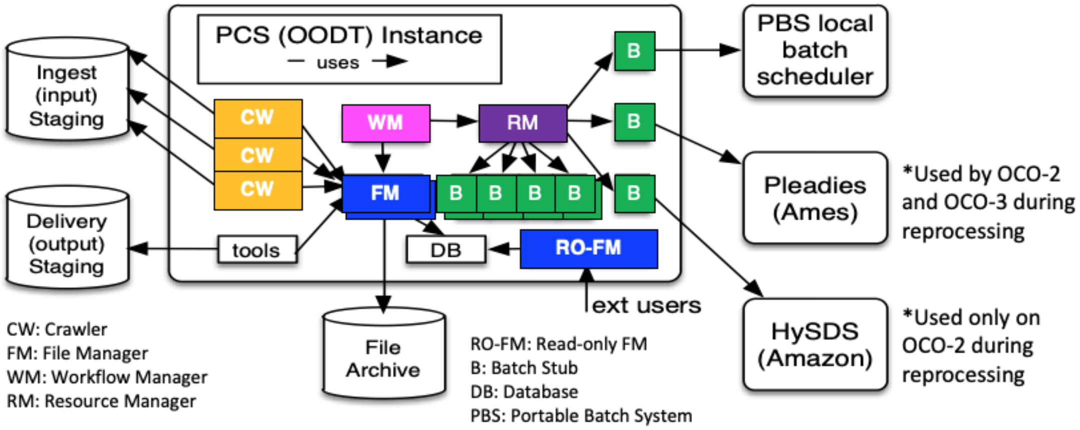
OCO-2/OCO-3

Orbiting Carbon Observatory (OCO)-2 was launched on July 2nd, 2014 and OCO-3 was launched on May 4th, 2019. OCO-3 was the flight spare of the OCO-2 instrument and was installed on ISS Japanese Experiment Module. The objectives of OCO are to retrieve estimates of the column-averaged dry air mole fraction of carbon dioxide (XCO2) at regional scales (>1000 km) for a period of 3 years and to demonstrate a precision of better than 0.25% [1 part per million or ppm]. Both OCO-2 and OCO-3 collect data in nadir, glint, and target modes, and OCO-3 has an additional snapshot area mapping mode.

The OCO science data system is depicted below:


JPL
External
Users (JPL and External)
NASA Ground Station*
Raw Data
Mission Operations System ᴼ
Ancillary Files
Global Modeling and Assimilation Office (GMAO)
Weather Model
Calibration Team
Calibration Parameters
All products
Science Data System
Receive/Ingest Data
• Instrument data
• Health and Status data
Store/Catalog Data Products
• Standard Data Products
• Ancillary/Auxiliary Data
• Internal & Temporary Data
Algorithm Development & Validation
• Develop algorithms & executables
• Host instrument calibration tools
• Host cal/val & analysis tools
Data Access
• Deliver to GES DISC
Data Processing
• Level 0, 1, 2 Processing
• Nominal processing
• Monthly reprocessing with updated calibration parameters
• Bulk reprocessing with updated calibration parameters and algorithms
Data Accountability & Quality
• Data Accountability
• Data Quality
• System Performance
• Product Latency
Public Users
Science Data Products
Telemetry and Data Products
Goddard Earth Sciences Data and Information Services (GES DISC)
All products
Project Users

OCO-2 and OCO-3 have separate data systems that leverage the same core architecture. The teams that work on OCO-2 and OCO-3 mission operations systems are different. OCO-2 telemetry comes from the Earth Orbiting System (EOS) Data and Operations System (EDOS), and OCO-3 telemetry comes from the Huntsville Operations Support Center (HOSC). Each project has 4 deployments of the Data Systems that are managed by the Operations Team. The first deployment supports a forward datastream, which ingests telemetry on a daily basis and processes with the most up-to-date calibration coefficients to produce results quickly and alert the team of any potential issues that might require immediate attention; the second supports a reprocessing stream, which can be used to reprocess the month’s data with updated calibration parameters and to reprocess all available mission data with updated algorithms and calibration parameters; and, the third and fourth support a science computing facility that supports testing and validating the algorithm with flexible configuration for the workflow and software. The system architecture is depicted below:



*Used by OCO-2 and OCO-3 during reprocessing

*Used only on OCO-2 during reprocessing

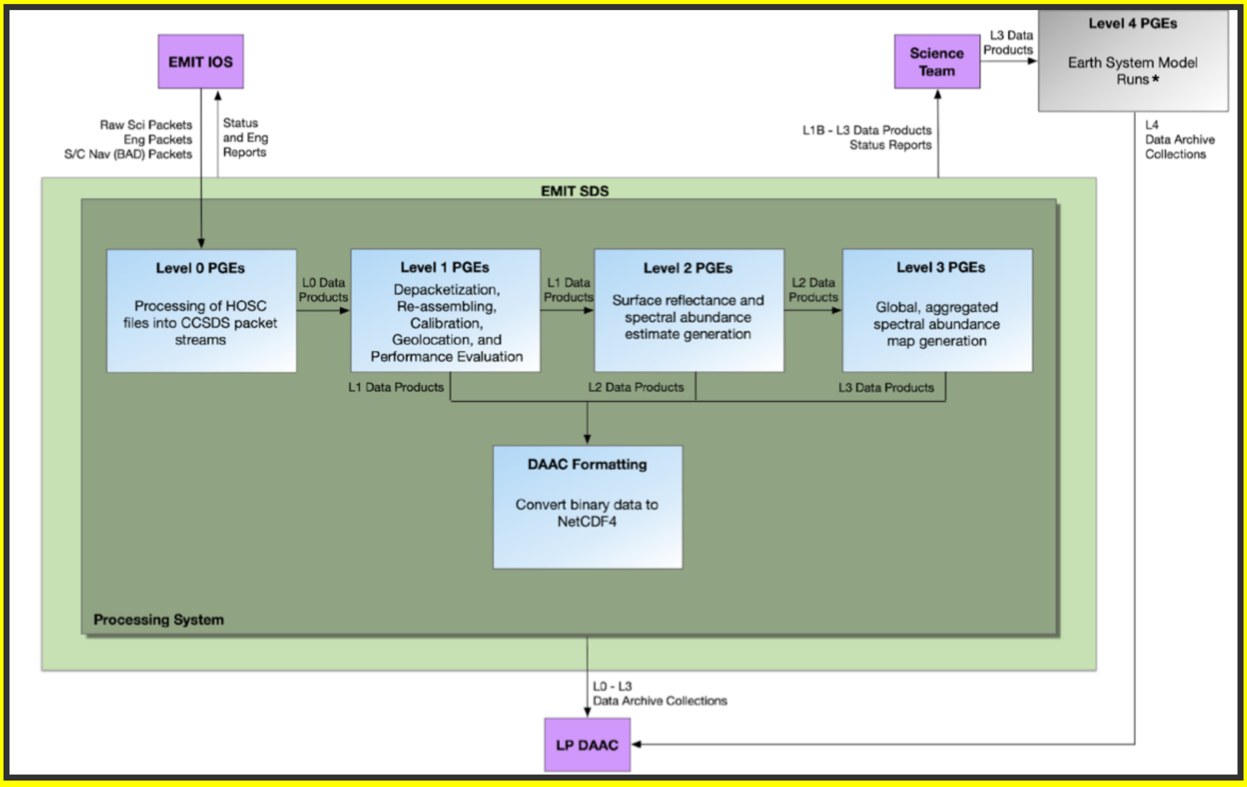
CW: Crawler
FM: File Manager
WM: Workflow Manager
RM: Resource Manager

RO-FM: Read-only FM
B: Batch Stub
DB: Database
PBS: Portable Batch System

EMIT

EMIT (Earth Surface Mineral Dust Source Investigation) is a comprehensive spectroscopic measurement of the Earth's mineral dust source regions to initialize state-of-the-art Earth System Models. EMIT will constrain the sign and magnitude of dust-related radiative forcing and predict the increase or decrease of available dust sources under future climate scenarios. It will be launched to the ISS in 2022.

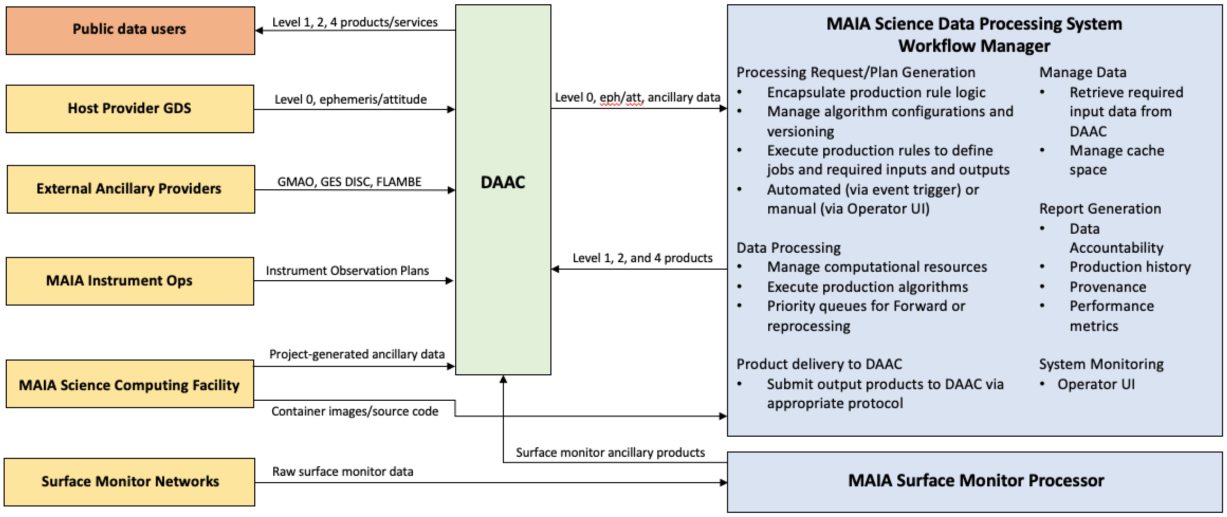
EMIT SDS will use the available JPL Imaging Spectrometer Computing Facility. Level 0 to Level 3 data will be processed at the JPL computing facility, while the Level 4 data will be produced by Earth system models elsewhere. The SDS overview is depicted below:



MAIA

The Multi-Angle Imager for Aerosols (MAIA) was selected under the Earth Venture Instrument-3 solicitation. Its primary science objective is to assess the impacts of particulate matter (PM) on adverse health outcomes. MAIA is currently in phase C with the likely launch date no earlier than 2024. MAIA observations are targeted rather than global. 'First Look' Level 2-4 data are processed using forecast meteorology data, and 'Final' Level 2-4 data are processed using reanalysis. MAIA processing is designed to execute different combinations of Program Generated Executables (PGEs) for each target and a given PGE can have a target-specific configuration. Information from the daily instrument observation plan and a target list will be used to construct a workflow for mission data processing.

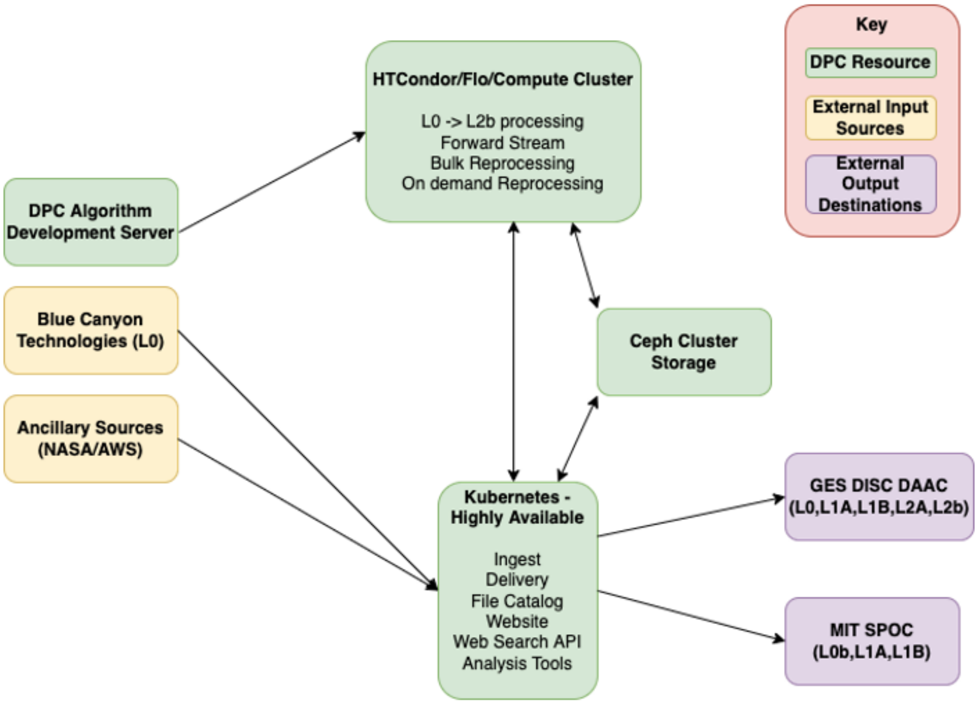
The data processing system architecture is depicted below:


Public data users
Level 1, 2, 4 products/services
Host Provider GDS
Level 0, ephemeris/attitude
External Ancillary Providers
GMAO, GES DISC, FLAMBE
MAIA Instrument Ops
Instrument Observation Plans
Project-generated ancillary data
MAIA Science Computing Facility
Container images/source code
Surface Monitor Networks
Raw surface monitor data
DAAC
Level 0, eph/att, ancillary data
Level 1, 2, and 4 products
Surface monitor ancillary products
MAIA Science Data Processing System Workflow Manager
Processing Request/Plan Generation
• Encapsulate production rule logic
• Manage algorithm configurations and versioning
• Execute production rules to define jobs and required inputs and outputs
• Automated (via event trigger) or manual (via Operator UI)
Data Processing
• Manage computational resources
• Execute production algorithms
• Priority queues for Forward or reprocessing
Product delivery to DAAC
• Submit output products to DAAC via appropriate protocol
Manage Data
• Retrieve required input data from DAAC
• Manage cache space
Report Generation
• Data Accountability
• Production history
• Provenance
• Performance metrics
System Monitoring
• Operator UI
MAIA Surface Monitor Processor

TROPICS

Time-Resolved Observations of Precipitation structure and storm Intensity with a Constellation of Smallsats (TROPICS) consists of 6 Cubesats in three low-Earth orbital planes. UW-Madison was chosen by the TROPICS project to be the data processing center (DPC). The DPC ingests Level 0 data and creates Level 1 and Level 2 products, including browse imagery using algorithms from TROPICS Science Team Members. The DPC delivers all data to the NASA GES DISC for archive, which is also available to the general public. The DPC supports the Science Team with access to data products, support in their development of algorithms and validation of data products through a development server and the TROPICS DPC website. The DPC follows the same delivery structure as the Atmosphere SIPS. Currently, the TROPICS DPC runs in on-prem cloud (Kubernetes) at UW-Madison.

The TROPICS DPC infrastructure is depicted below:



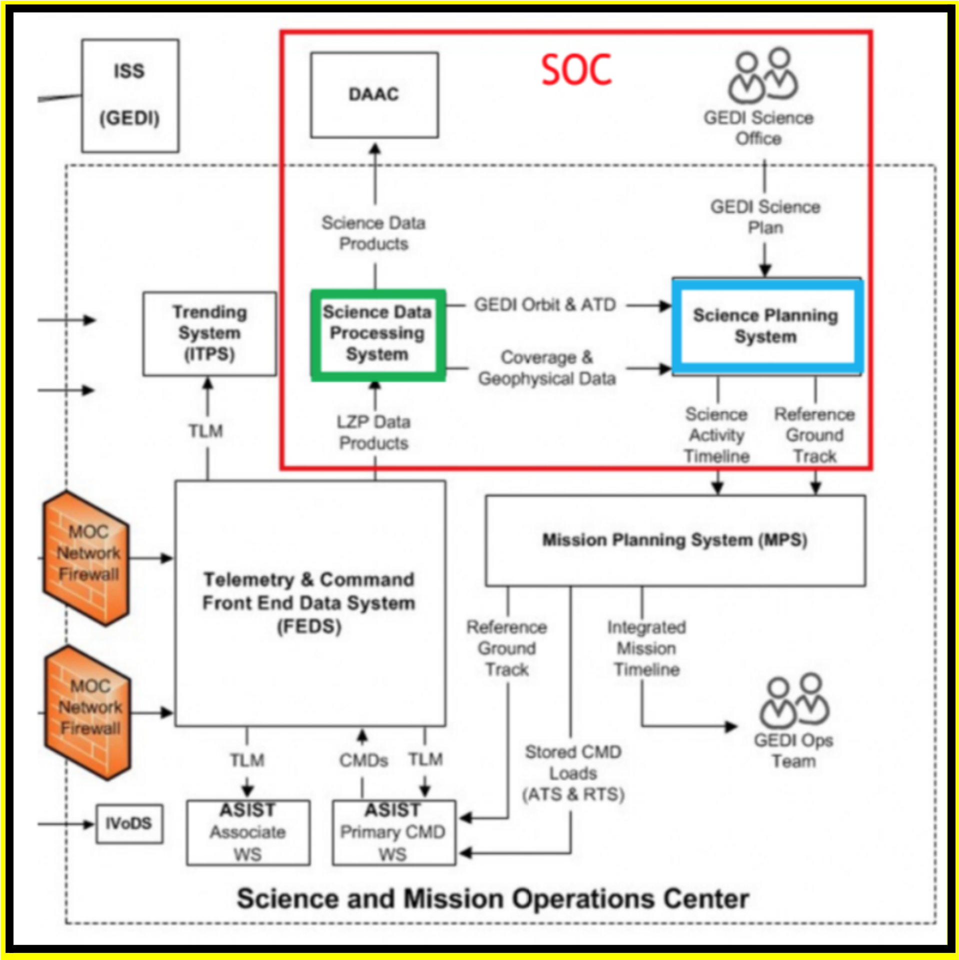
DPC leverages experience and shared resources from other projects which has proven to be an efficient and cost-effective method. It supports the NASA-funded Science Team per the requirements of the TROPICS project and focuses on creating unique tools or incorporates existing tools for Science Team members to more accurately analyze and improve their products.

GEDI

GEDI (Global Ecosystem Dynamics Investigation) is a NASA Earth Ventures Instrument aimed to advance our ability to characterize the effects of changing climate and land use on ecosystem structure and dynamics. GEDI was launched in December 2018 and deployed on the ISS JEM-EF, an external platform for conducting scientific observations, Earth observations, and experiments in an environment exposed to space.

GEDI Science operations Center (SOC) is composed of a science planning system (SPS) and a science data processing system (SDPS). The main functions of SPS are receiving Science Plan from GEDI Science Office, predicting positioning and pointing to compute optimal ground track sampling, producing Science Activity Timeline (SAT) and Reference Ground Track (RGT), which are provided to Mission Operations Center (MOC) for command load. The main functions of SDPS are receiving Level 0B data from MOC, generating Level 1 – 3 Science Data Products, distributing data products to DAACs. The Levels 0 to 2 data go to the LP.DAAC and the Levels 3 and 4 go to the ORNL DAAC.

An overview of the SOC is depicted below:



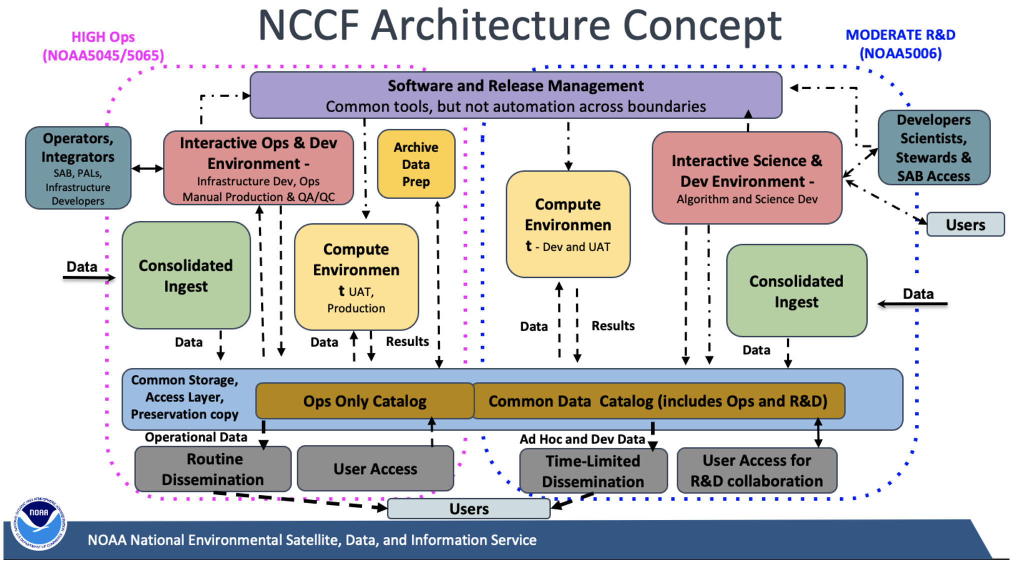
7.2.4 Non-NASA Missions

National Oceanic and Atmospheric Administration (NOAA)

The National Environmental Satellite, Data, and Information Service (NESDIS) was created by NOAA to operate and manage the United States environmental satellite programs, and manage the data gathered by the National Weather Service and other government agencies and departments. Legacy NESDIS Ground Enterprise (NGE) is a disparate set of systems with an unsustainable approach to accommodate the growing volume of observing system data. The NESDIS Common Cloud Framework (NCCF) will transform to a new approach to accommodate these growing data volumes.

NCCF provides end-to-end ground capabilities with a cloud agnostic common enterprise architecture. It will use industry standard, scalable, and open source (where possible) cloud service provider (CSP) ubiquitous managed services, leveraging Docker for science-driven containerized cloud algorithm packages (CCAPs) for product processing on Linux-based virtual machines, using a centralized GitLab service for CI/CD orchestration. NESDIS Cloud-sandbox Infrastructure Services (NCIS) provides a research and development component for piloting and prototyping. The National Centers for Environmental Information (NCEI) is piloting a new data archive and data access services under the NESDIS Cloud Archive Program (NCAP). It is based on serverless cloud-based services and aims to be data catalog agnostic, scalable, reliable, and highly available. It has already integrated NASA's Common Metadata Repository (CMR) for enterprise testing.

The NCCF Architecture is depicted below:


NCCF Architecture Concept
HIGH Ops (NOAA5045/5065)
MODERATE R&D (NOAA5006)
Software and Release Management
Common tools, but not automation across boundaries
Operators, Integrators
SAB, PALs, Infrastructure Developers
Interactive Ops & Dev Environment -
Infrastructure Dev, Ops Manual Production & QA/QC
Archive Data Prep
Compute Environment - Dev and UAT
Interactive Science & Dev Environment -
Algorithm and Science Dev
Developers Scientists, Stewards & SAB Access
Users
Data
Consolidated Ingest
Compute Environment UAT, Production
Data
Results
Data
Results
Consolidated Ingest
Data
Data
Data
Common Storage, Access Layer, Preservation copy
Ops Only Catalog
Common Data Catalog (includes Ops and R&D)
Operational Data
Ad Hoc and Dev Data
Routine Dissemination
User Access
Time-Limited Dissemination
User Access for R&D collaboration
Users
NOAA National Environmental Satellite, Data, and Information Service

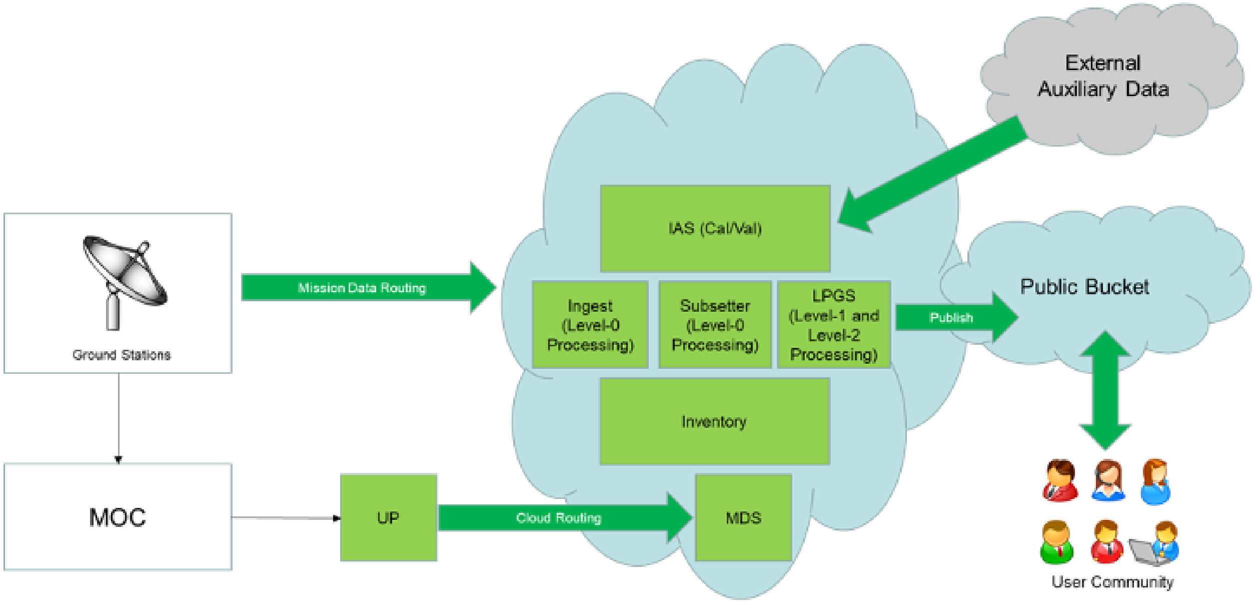
The NCCF design is based on a multi-account and multi-Virtual Private Clouds (VPC) architecture to run specific functions, with a separate/dedicated management and IT Security VPC. All NCCF infrastructure is implemented with IaC (Infrastructure as Code) and stored in a code repository. Its modular approach allows scalability to add additional services as needed.

US Geological Survey (USGS) (Landsat Mission Data Processing System)

The USGS is migrating the existing on-premises processing into the Landsat Cloud (LC) which will produce Level 1 - 3 data and Analysis Ready Data (ARD) tiles. The LC is built using AWS services (where appropriate), including S3 cloud object storage, Kubernetes container orchestration services, Aurora relational database, Simple Queue Service (SQS), and Lambda functions for event-driven serverless computing. The system development approach uses open source packages / technologies where possible and relies on vendor-specific capabilities where they make sense. All Landsat algorithms are publicly available.

The Landsat processing and distribution system is rapidly adopting standards to provide for more modern and consistent access to data products, including SpatioTemporal Asset Catalog (STAC) metadata records, Cloud-Optimized GeoTIFF (COG) extensions in data products and using XML-based metadata to augment existing legacy metadata.

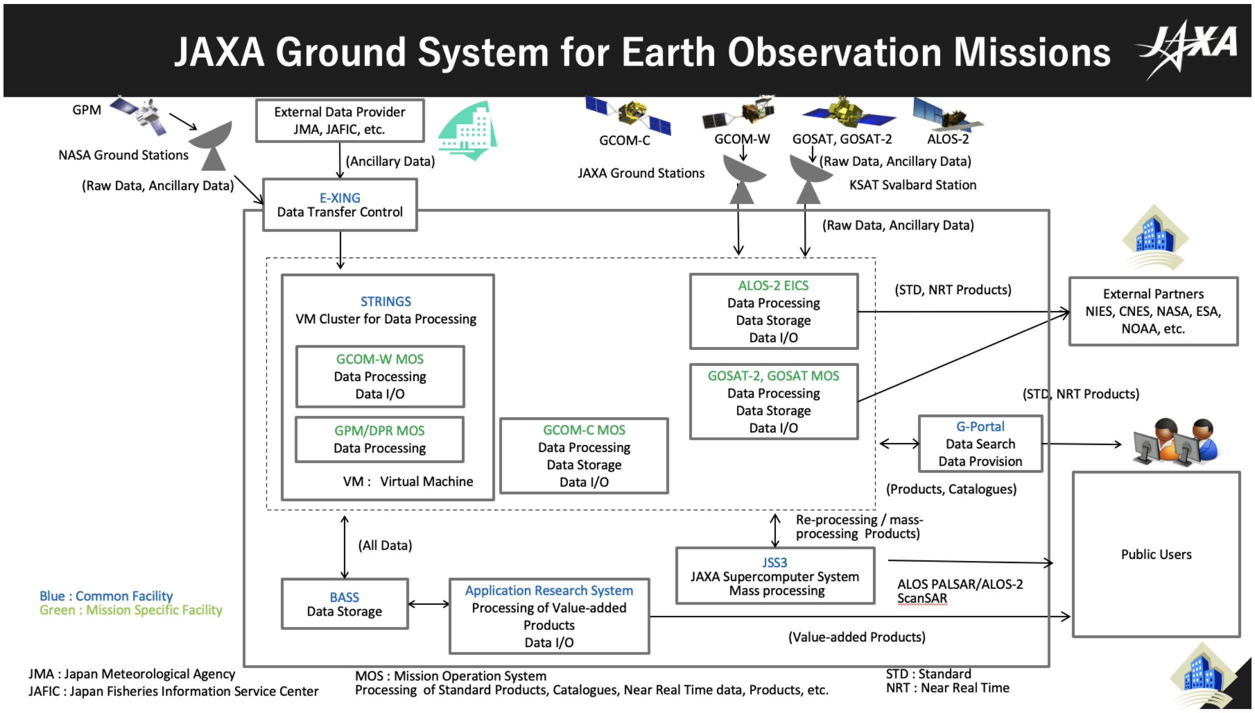
The Data Processing System Architecture is depicted below:



Japan Aerospace Exploration Agency (JAXA)

Many of JAXA's individual project data portals are aggregated into a single Earth-graphy web portal (https://earth.jaxa.jp/en) that provisions data to the G-Portal, a single point for mission standard product data dissemination, and a number of other thematic and partner portals. JAXA is also a partner in the Earth Observing Dashboard (https://eodashboard.org) and in cooperation with Google Earth Engine to enable processing of Analysis Ready Data (ARD). Additionally, a new Japanese national satellite data platform "Tellus" targets enhancement of satellite data utilization for business purposes, providing data, AI, and software free of charge. It provides user-friendly tools and software to be used with ARD and computing resources in the Cloud.

The JAXA data processing system is depicted below including mission-specific facility (green) and common-facility (blue):



JAXA Ground System for Earth Observation Missions

JMA : Japan Meteorological Agency
JAFIC : Japan Fisheries Information Service Center
MOS : Mission Operation System
Processing of Standard Products, Catalogues, Near Real Time data, Products, etc.
STD : Standard
NRT : Near Real Time

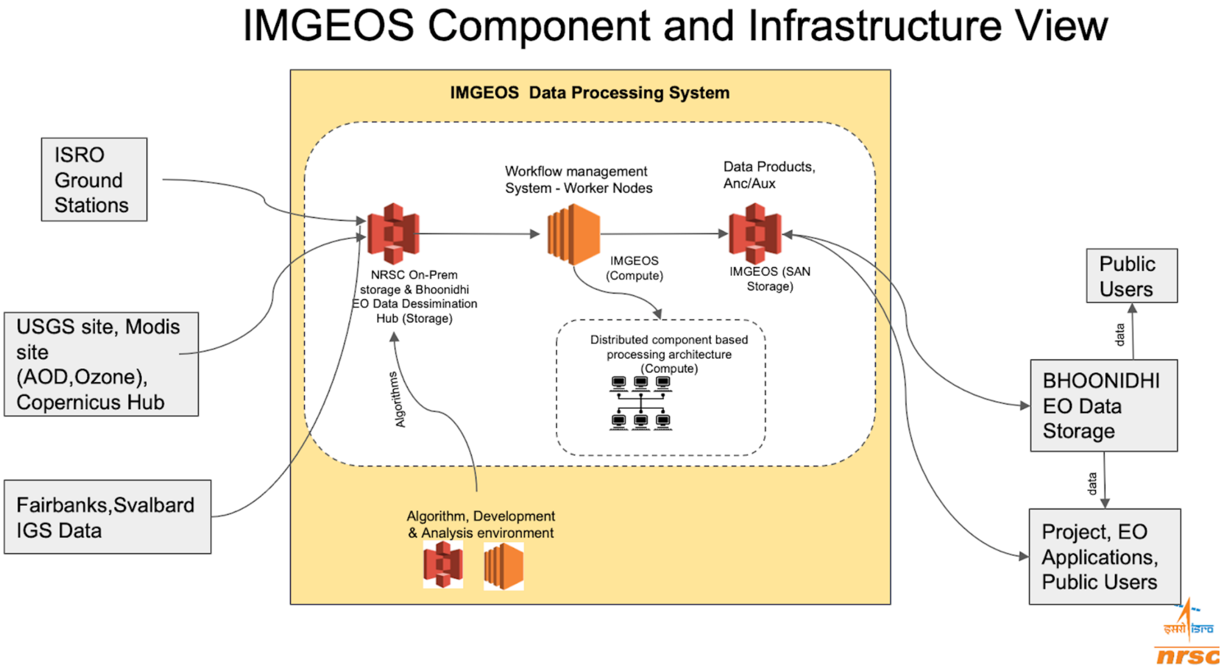
Public availability of JAXA data depends on the resolution of the data. Mid to low resolution data, documentation, libraries, tools, and sample programs are open and freely available to the public. High Resolution Data, some processing and analysis tools and calibration and validation data are protected or licensed. While there remain some challenges regarding Intellectual Property (IP), JAXA is working to support Open-Source Science to promote scientific and application research for the next generation.

National Remote Sensing Center, India Space Research Organization (ISRO)

The Integrated Multi-Mission Ground Segment for Earth Observing Satellites (IMGEOS) provides multi-mission data processing of 95 passes/day or about 1.5 TB/day. Through 24/7 operations, it supports algorithm development and analysis. The infrastructure is based on Open Source technologies including Linux, GNU, Python, Open JDK, PostgreSQL, MySQL, Maria DB, and Open Virtualization Format (OVF).

Data dissemination includes both open and priced products, available through a centralized portal (https://bhoonidhi.nrsc.gov.in) deployed with Cloud based technologies, offering data from 40 satellites, with 30 years of data. The portal has an accessible on demand processing platform enabling users to develop and test user defined processing workflows in the cloud.

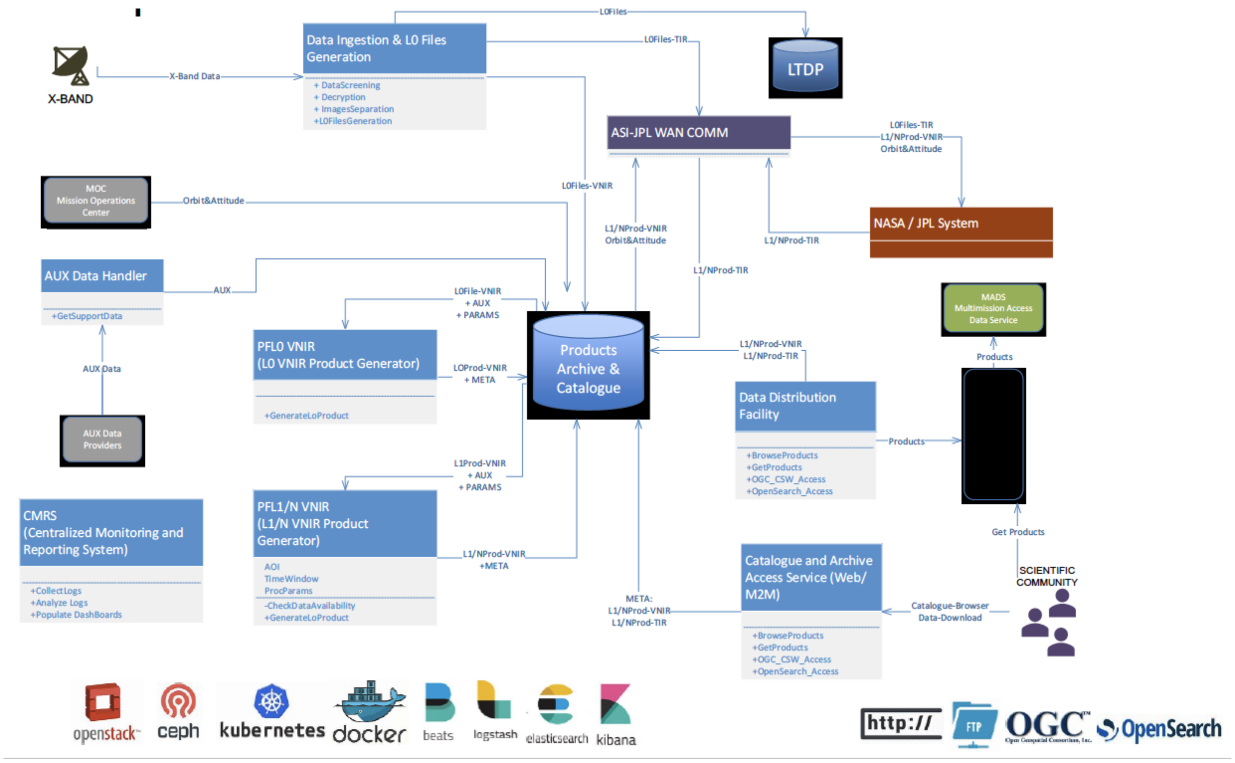
The IMGEOS architecture is depicted below:


IMGEOS Component and Infrastructure View
IMGEOS Data Processing System
ISRO Ground Stations
USGS site, Modis site (AOD,Ozone), Copernicus Hub
Fairbanks,Svalbard IGS Data
Workflow management System - Worker Nodes
Data Products, Anc/Aux
NRSC On-Prem storage & Bhoonidhi EO Data Dessimination Hub (Storage)
IMGEOS (Compute)
IMGEOS (SAN Storage)
Distributed component based processing architecture (Compute)
Algorithms
Algorithm, Development & Analysis environment
Public Users
data
BHOONIDHI EO Data Storage
data
Project, EO Applications, Public Users
इसरो isro
nrsc

Italian Space Agency (ASI)

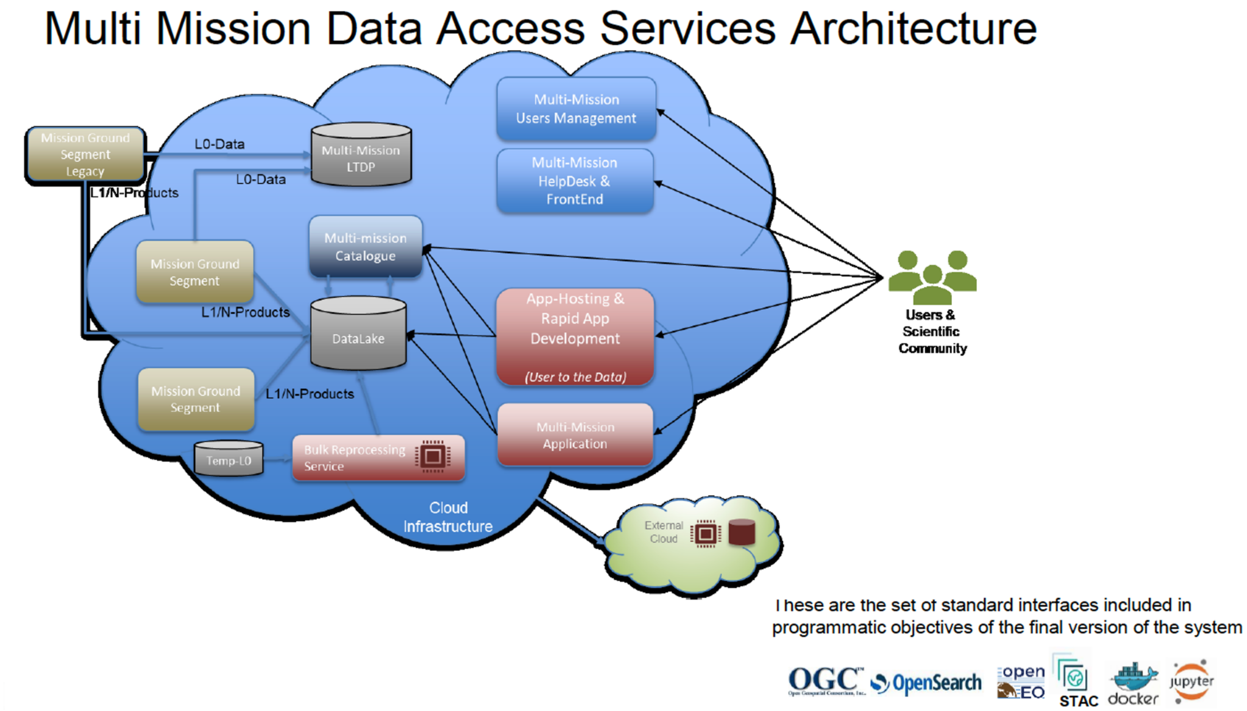
The Surface Biology Geology (SBG)-HEAT ITA Scientific Operation Center (SOC) processes data and sends products to the ASI Mission Access Data System (MADS). MADS provides a multi-mission data archive and value-added services, including bulk-reprocessing. MADS exploits the elasticity of the cloud infrastructure to get new resources, locally or from an external cloud, to install new applications that can use the multi-mission data in the archive. MADS can provide bulk reprocessing, possibly running on an external cloud-infrastructure.

The architectures for the SOC is depicted below:



MADS supports open science by allowing users to directly execute code on the same infrastructure where the data are stored (user-to-the-data) without the need to transfer it. Users can upload their own applications using Jupyter Notebooks or run in a docker container. Users can easily share the results of their applications and algorithms with other users. Some missions require formal acceptance of a user license with precise restrictions on the distribution and exploitation of data.

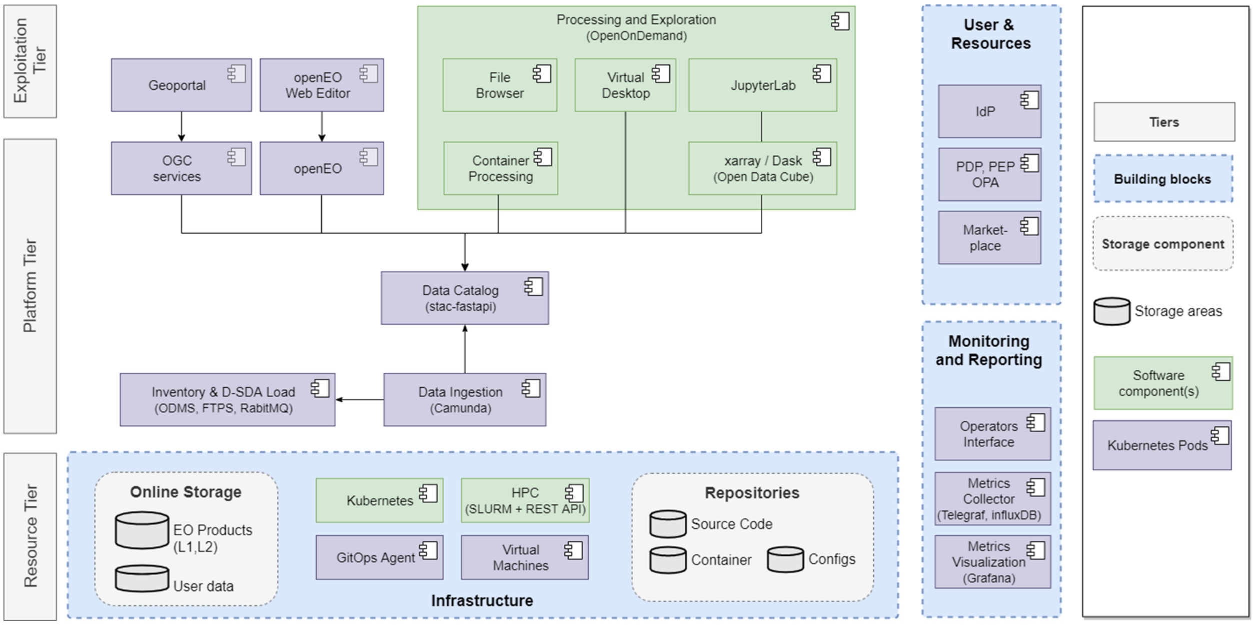
The MADS architecture is depicted below:


Multi Mission Data Access Services Architecture
Mission Ground Segment Legacy
L0-Data
L0-Data
L1/N-Products
Multi-Mission LTDP
Multi-Mission Users Management
Multi-Mission HelpDesk & FrontEnd
Mission Ground Segment
Multi-mission Catalogue
L1/N-Products
DataLake
App-Hosting & Rapid App Development
(User to the Data)
Mission Ground Segment
L1/N-Products
Multi-Mission Application
Temp-L0
Bulk Reprocessing Service
Cloud Infrastructure
External Cloud
Users & Scientific Community
These are the set of standard interfaces included in programmatic objectives of the final version of the system
OGC
OpenSearch
openEO
STAC
docker
jupyter

German Aerospace Center (DLR)

The “Large-Scale Data Mining in Earth Observation” project is using deep/machine learning in Earth observation labeling. The DLR “Terrabyte” platform makes Earth observation data accessible for research and offers practical tools for analytics. It connects the DLR satellite data with managed online storage of around 100 PB using the supercomputers of the Leibniz Supercomputing Centre (LRZ). Terrabyte is a viable alternative to commercial data clouds meeting security and data protection requirements, and addressing the co-location of data and compute for “Big Geo-Data processing.” It can support collaborative science projects and experiments in a hybrid HPC-cloud environment.

The architecture for the High-Level Data Processing System is depicted below:

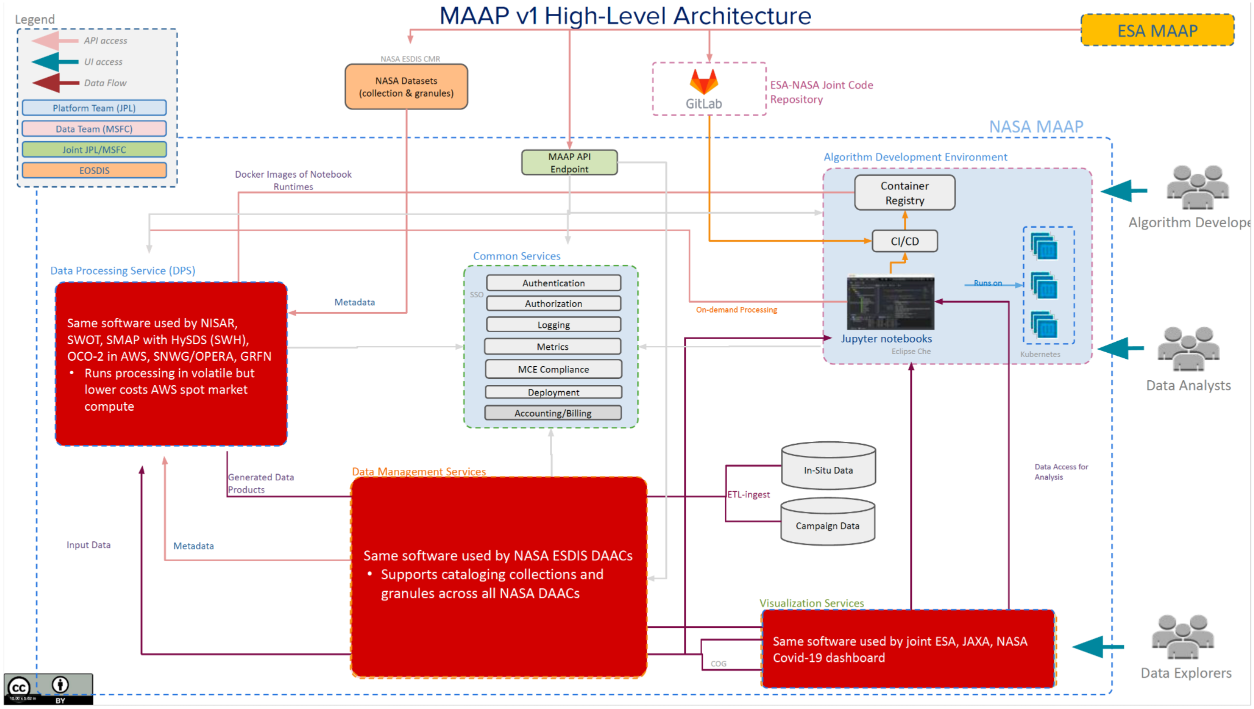


Terrabyte hosts a mix of public and private datasets, and offers a *hot* data cube extension that formats geospatial data cubes for machine learning and AI community challenges as well as other research. This data can be fed directly into Jupyter Notebooks and browser-based code-development interfaces as well as directly to the Supercomputers. They foster a collaborative culture through open sharing and support of the FAIR principles (Wilkinson et al., 2016).

European Space Agency (ESA)

The Multi-Mission Algorithm and Analysis Platform (MAAP) is a joint NASA-ESA project to develop a virtual open and collaborative environment accessible via a web browser that brings together EO data and in-situ data with computing resources and hosted processing. It provides collaborative tools and forums for users to exchange experiences.

The MAAP v1 High-Level Architecture is depicted below:



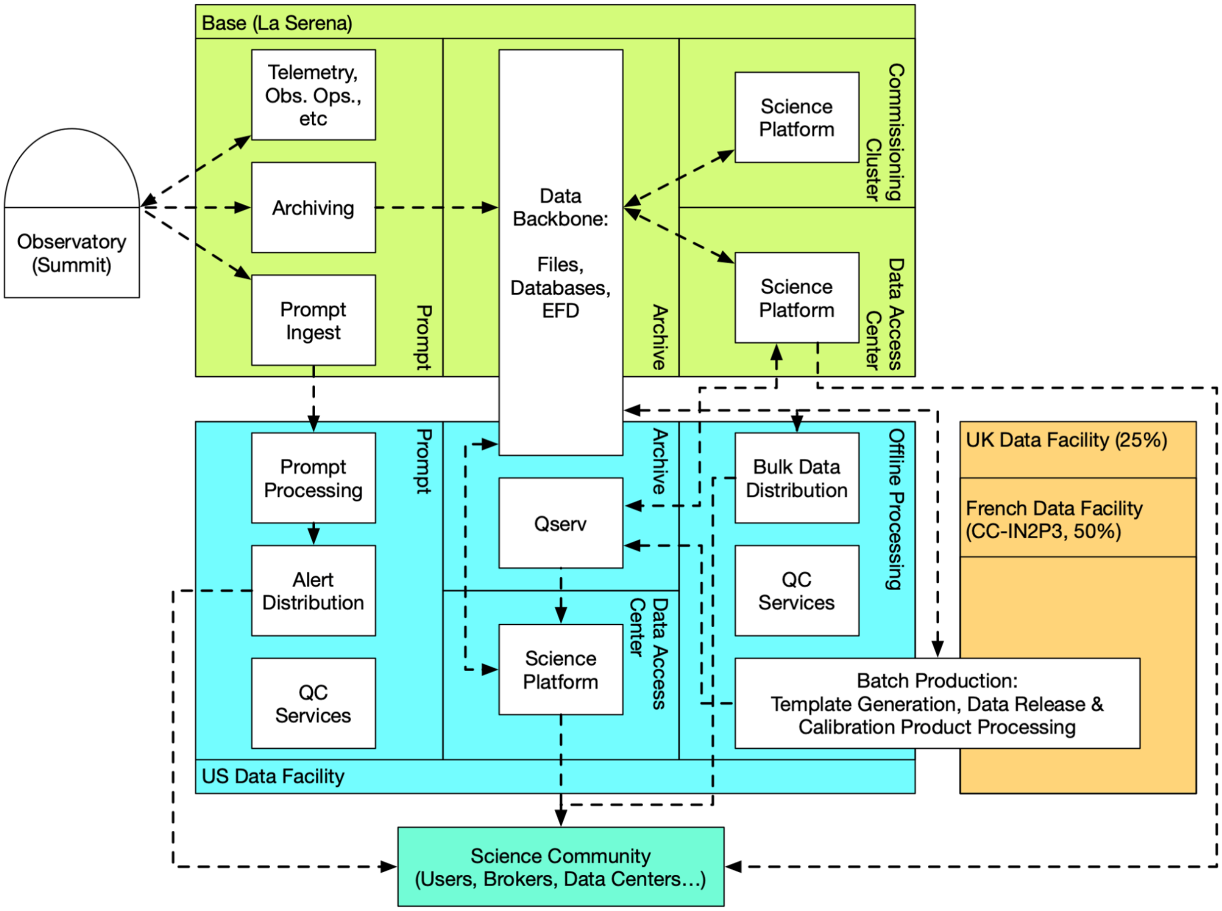
MAAP supports a Product Algorithm Laboratory (PAL) approach that makes development and implementation of processing algorithms easier and faster to mature. It supports Open Science allowing people outside the core science team to contribute to the product improvement cycle.

7.2.5 Non-Earth Missions

Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST)

The LSST (Ivezić et al., 2019) has four main science themes: probing dark energy and dark matter, taking an inventory of the solar system, exploring the transient optical sky, and mapping the Milky Way. LSST is currently in Construction, Pre-Operations, and Commissioning; Operations to start in early 2024. The data management system will process 20TB/night of raw data. From 60 PB of raw data, it will produce up to 500 PB of data products that will be public after 2 years. The system will run in a hybrid model using on-premises compute for most processing with public cloud data access centers.

The LSST Data Management Architecture is depicted below:

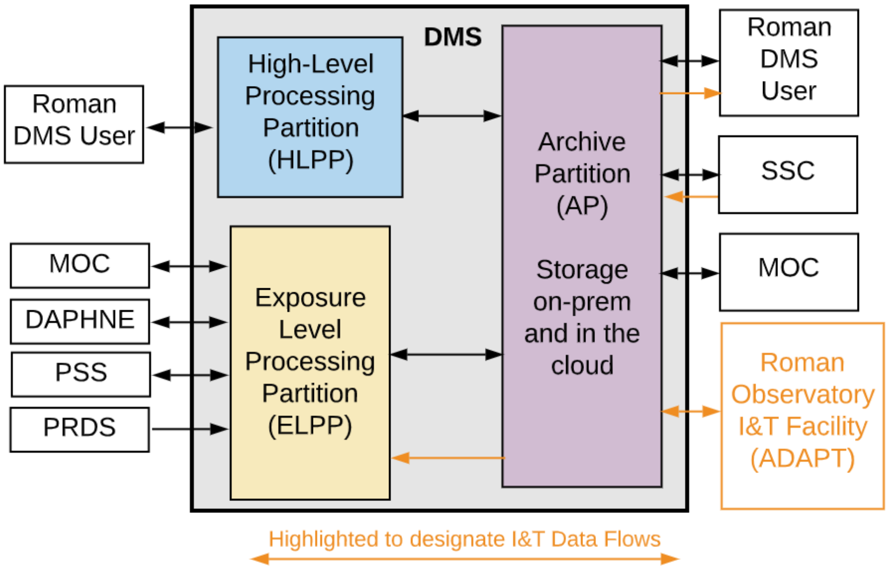


In support of open science, all code will be published in public repositories on GitHub. A public process will determine best-of-breed algorithms/services in key areas. Community-provided modifications or new algorithms can be incorporated into the next year's Data Release or the Alert Production. All binary artifacts, including container images will be public as well. Some infrastructure or internal systems may use proprietary products.

Nancy Grace Roman Space Telescope

The Roman Space Telescope is a NASA observatory designed to unravel the secrets of dark energy and dark matter, search for and image exoplanets, and explore many topics in infrared astrophysics. It is currently scheduled to launch in October, 2026. The Data Management System (DMS) exists as a hybrid architecture of on-premises and cloud processing environments and archive storage. It will receive, process, and archive up to 13.9 Terabits of compressed science telemetry per day, archiving 18.2 PB of science data products over the mission, distributing 97 PB of data per year by year 5. All data is non-proprietary and comes with no exclusive access limitations. High Level Processing runs in the STSci AWS cloud, using cloud native and compatible tools and frameworks, including S3 for cloud archive storage. The science calibration software is open-source on GitHub. GitLab is used internally as well as docker/containers, and kubernetes.

A diagram depicting the DMS architecture is shown below:


DMS
High-Level Processing Partition (HLPP)
Archive Partition (AP)
Storage on-prem and in the cloud
Exposure Level Processing Partition (ELPP)
Roman DMS User
MOC
DAPHNE
PSS
PRDS
Roman DMS User
SSC
MOC
Roman Observatory I&T Facility (ADAPT)
Highlighted to designate I&T Data Flows

For open science, a 'Roman Science Platform' is also provided in the cloud that can use JupyterHub, giving scientists easy to access analysis options that don't require data download, along with a managed environment.

IPAC Missions

The speaker was not available.

7.2.6 System Interfaces

Earth Science Data Information System (ESDIS)

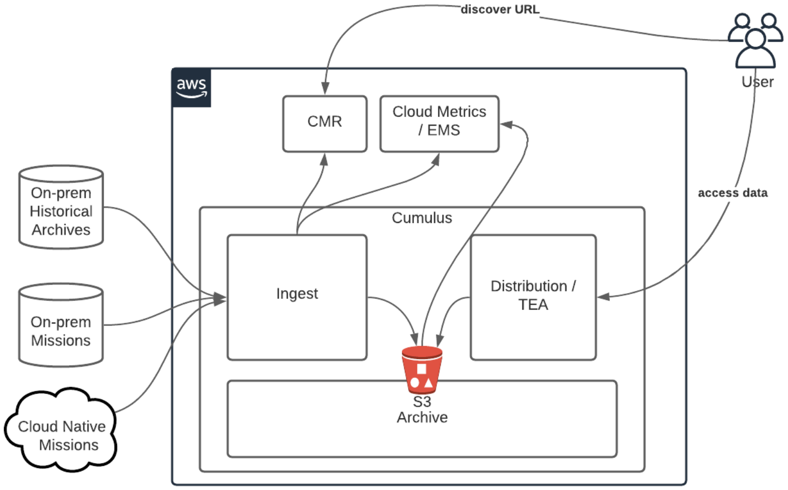
The Earth Science Data Information System (ESDIS) project manages the EOSDIS ensuring that the science data systems provide scientific data stewardship for all data collections; provide a unified and simplified environment for access by diverse and distributed communities; evolve, grow, and adapt new data to new data system technologies; expand and engage with the user community to improve/enhance data access; and partner with other organizations, agencies, and international partners to share data. The Earth Observing System Data and Information System (EOSDIS) is a comprehensive information system designed to support and manage NASA's Earth Observing (EO) mission data. It is a distributed system of data archives, processing systems, and networks designed to ingest, archive, distribute, and visualize satellite observations and Earth Science data, which include field campaign measurements, airborne data, in situ data, model data and ancillary products used for processing. The EOSDIS archives and distributes data through the Data Active Archive Centers (DAACs) each designed to serve a specific user community: Atmosphere, Ocean, Cryosphere, Land, and Human Dimensions. In 2021, EOSDIS distributed nearly 61 PB of data and a total of 1.9 billion files to end users.

EOSDIS adds contributions through an interactive process involving NASA Headquarters (HQ), Science Investigator-led Processing Systems (SIPS), and Distributed Active Archive Centers (DAAC) with a significant amount of time spent in Quality Assurance (QA) of data. EOSDIS interfaces with the Mission Systems who receive data downlinked from the spacecraft to the ground stations, which is processed into Level 0 products and distributed to the SIPS. The SIPS generates science mission products for the Earth Observing System (EOS including Suomi National Polar-orbiting Partnership (SNPP) and the Joint Polar Satellite System (JPSS) missions) and delivers them to the DAACs. The SIPS are managed by Principal Investigators (PI) and Science Teams. There are 11 SIPS and not all of the SIPS do processing. Some of those SIPS deliver algorithms to a DAAC to do the processing (e.g. AIRS). The PI and algorithms for standard products are competitively selected through a NASA HQ Science Program Award. The PIs develop the Program Generation Executable (PGE), which are integrated into the designated discipline-based SIPS for (re-)processing of mission standard products. These products along with preservation items (e.g., Algorithm Theoretical Basis Documents - ATBDs) are delivered to the DAAC for long-term archive. The DAACs serve as the primary point of contact for mission science teams and are responsible for the data and information management, quality of distributed products, and support for open-source software and cross-mission science and modeling.

The DAACS are discipline-oriented and have expertise for managing data for specific communities based on years of working with PIs and Science Teams. Discipline-oriented DAACS cater to variations in: 1) how data is used (e.g., different data formats and access needs); 2) how to connect with and understand a specific user community's needs (e.g., tools and services) through active User Working Groups with representatives from those disciplines focused on advising NASA; and 4) how to work with Earth Science Directorate (ESD) program scientists to plan for developing data collections, addressing science topics, and serving as a link to the community. Missions are assigned to DAACS based on discipline. The new Earth System Observatory (ESO) missions are multi disciplined, and DAACS will need to be able to support multiple interdisciplinary science.

EOSDIS includes many enterprise tools. Earthdata is the EOSDIS website that provides visibility to the interdisciplinary use of data and demonstrates how to use the data. The Common Metadata Repository (CMR) and Earthdata Search Client provides high performance data search and discovery across EOSDIS holdings. The Global Imagery Browse Services (GIBS)/Worldview and Giovanni services provide quick access to satellite imagery and data visualization tools to explore imagery covering every part of the world. Land Atmosphere Near real-time Capability for EOS (LANCE) supports users interested in monitoring natural and man-made phenomena using near real-time (NRT) data and imagery that is made available within 3 hours from satellite observation. The EOSDIS Metrics System (EMS) collects and reports on data ingest, archive, and distribution metrics across EOSDIS. The Earthdata Infrastructure (EDI DevOps) is a platform for requirement management, code development, testing, and deployment to operations. User support tools provide user relationship management and issue resolution. Earthdata Log-in provides a centralized and simplified mechanism for user registration and account management for all EOSDIS components. Finally, the Earthdata Forum is an interactive platform that allows subject matter experts to respond to users' science data questions.

The Earthdata Cloud ingest/distribution process is shown below.



With the steady increase in data distribution (e.g., NASA will produce 50 PB of Synthetic Aperture Radar (SAR) data each year), compute, and storage needs, instead of continuing the traditional paradigm of distributed systems located across all the US, EOSDIS is moving to the cloud, Earthdata cloud (EDC). The EDC became operational in July 2019 and is a managed commercial cloud architecture. The EDC will improve efficiency of NASA's data system operations and will maintain a free and open data policy. With transitioning to the cloud, EOSDIS is able to realize several end user benefits. These include users being able to: access processing power next to the data; improved performance; reduced time to move, manage, and store large volumes of data; and co-location of data where users can easily work with multiple EOSDIS datasets together with the option to download the data, if they prefer.

NASA Transform to Open Science

NASA's Open-Source Science Transform to Open Science (TOPS) initiative is a $40 million dollar 5-year program, across all 5 NASA Science Divisions, with the objective to increase understanding and adoption of open science, accelerate major scientific discoveries, and broaden participation by historically underrepresented communities. Open-Source Science (OSS) increases participation by more people with diverse experiences and allows the sharing of hidden knowledge that enables new applications of data and science while accelerating the pace and impact of science.

The OSS approach activates Open Science through transparency, accessibility, reproducibility, and inclusivity from the very beginning of the process. In transparent open science, all science processes and results are visible, accessible, and understandable by all user communities. With accessible open science, all data, tools, software, documentation, and publications are FAIR (Wilkinson et al., 2016) and diverse groups are welcomed to participate and collaborate in science. Lastly, reproducible, open science is where the scientific process and results are reproducible by the community and not just by the scientific community.

NASA's plan is for 2023 to be the Year of Open Science (YOOS). TOPS is energizing and uplifting OSS across the scientific communities through various avenues to promote visibility (e.g. through social media, conferences, articles/announcements), capacity sharing, and incentives for moving towards open science. Through an OSS free on-line course, organized as a scientific workflow, it will offer resources for teaching and advancing skills within the community to achieve open science, to convey the benefits to the greater scientific community, learning how to use open science tools, how to effectively use and share data, and best practices for sharing data. Teams will receive the Open Science Badge once they've completed the five modules (See Science Collaboration Approaches for more information). Additionally, incentives such as high-profile awards to reward significant leadership and progress toward open science showcasing the benefits and evaluating and updating existing awards and recognitions to include open science activities as review criteria. Lastly, the proposed plan is to use 2023 YOOS to announce sweeping changes across funding decisions, awards, promotions, evaluations, and the recognition of teams as well as individuals. YOOS will initiate the following conditions: require open FAIR data, open software and access to publications, pathway to earning open science badges, funding decisions will consider open science activities as part of review criteria, awards, promotions, and evaluations will consider OSS activities with the goal to have 90% of NASA science to have a Open Science Badge by 2027.

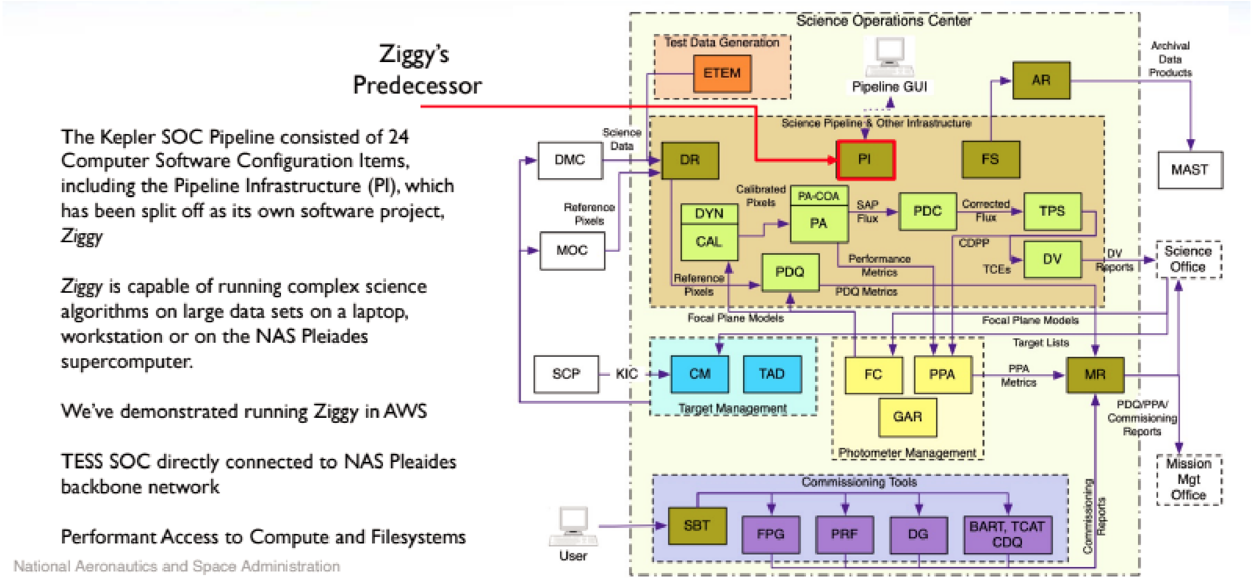
NASA's High-End Computing Capability (HECC): Growing to Support Science Data Processing for the Earth System Observatory Missions

NASA's only private High Performance Computing (HPC) cloud infrastructure has similar economies of scale to Cloud Service Providers (CSPs) who run multiple hyperscale data centers and a robust infrastructure that can easily be expanded. The HECC cost is a predictable fixed budget that is all inclusive of facilities, power, personnel (including data analysis and visualization services), hardware, networking, maintenance, and storage. The service is free to the consumer and the costs are covered by the Science Mission Directorate (SMD).

In order to meet tomorrow's computational challenges (e.g. SBG and NISAR missions), NASA's High-End Computing Capability (HECC) has evolved to support hybrid computing. HECC has been making enhancements by: tailoring file systems for large observational data sets (e.g. many small files with random file I/O); factoring compute node needs of observational dataset requirements into hardware selections (e.g. balancing amount of memory to number of cores); enabling node scheduling and autoscaling to support different use cases such as reserving nodes to run processing, procure dedicated resources, and setup special queues for high priority work; and ensuring high availability of compute modules (e.g. fault tolerant networks).

The Hybrid Computing and Storage Architectures (HCSA) is being developed to meet agency needs. It is a large-scale architecture with lower processing costs relative to public cloud for sustained computing capabilities such as simulations, modeling, and Machine Language (ML) training. There are also connections between public and private clouds to optimize resource allocation and to leverage features of public cloud that complement private cloud. It offers low-cost storage (e.g. Wasabi) and is lower cost to users who can't afford the compute resources on a public cloud.

The group has built and operated several major science pipelines for missions including the Kepler mission and Transiting Exoplanet Survey Satellite (TESS). The Kepler mission had data rates of 1GB/day, and after 2 years, it took 10 months to reprocess all the mission data. In 2011, porting everything over to Pleiades at NASA AMES Research Center enabled faster data processing. The TESS mission data rates increased 26 times more data per day than the Kepler mission and was processing at much faster rates. They predicted the Kepler mission would take 23 days to process 1 month of data, but today TESS only takes 5 days.



Ziggy's Predecessor

The Kepler SOC Pipeline consisted of 24 Computer Software Configuration Items, including the Pipeline Infrastructure (PI), which has been split off as its own software project, *Ziggy*

Ziggy is capable of running complex science algorithms on large data sets on a laptop, workstation or on the NAS Pleiades supercomputer.

We've demonstrated running Ziggy in AWS

TESS SOC directly connected to NAS Pleaides backbone network

Performant Access to Compute and Filesystems

National Aeronautics and Space Administration

The current MDPS architecture that runs the pipeline is the Kepler Science Operations Center (SOC). The SOC runs Ziggy as a stand-alone software project for pipeline processing agnostic to the kind of science data it processes and the platform (e.g. workstation, laptop, NASA Advanced Supercomputing (NASA)) it runs on. The pipeline provides automated processing of large volumes of data, data accountability, and dispatching of large tasks to Pleiades, a NASA HPC. Ziggy is at Technology Readiness Level (TRL) 7, a Class C software, and is currently being released under NASA Open Software Initiative.

OGC Data System Standards

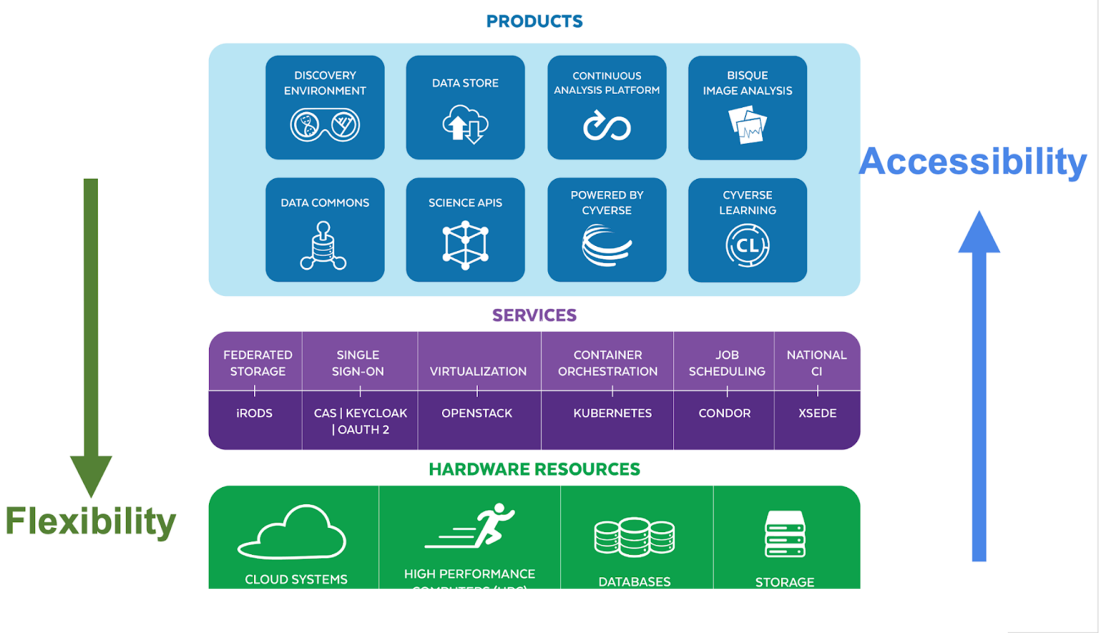
The speaker was not available.

7.2.7 Non-Mission MDPS

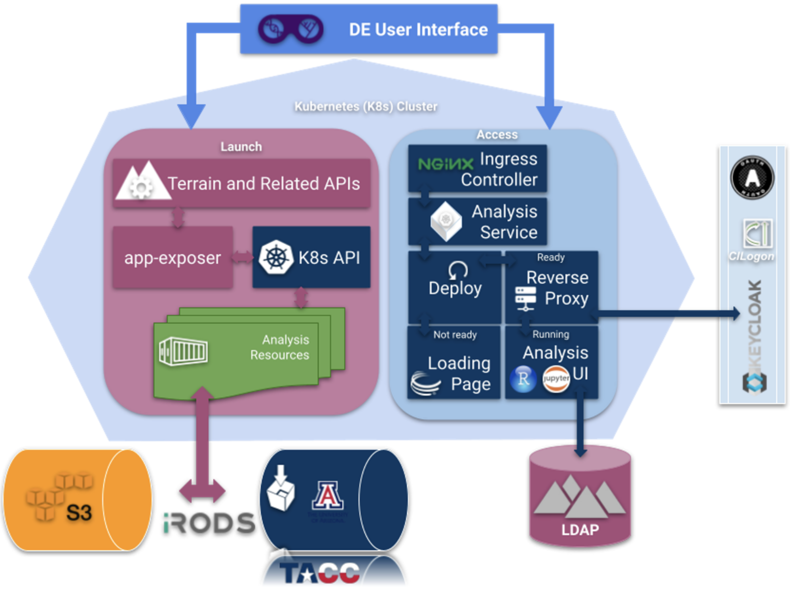
CyVerse: cyberinfrastructure for data driven discovery

CyVerse is a long-running National Science Foundation (NSF) project centered at the University of Arizona which began in 2008, initially focused on plant biology (iPlant), then transitioned to a more inclusive cyberinfrastructure in 2016, aiming to transform science through data-driven discovery. The platform has expanded its mission to serve users in the earth sciences, ocean sciences, astronomy, and social sciences. CyVerse is a cyberinfrastructure for data-driven discovery with a core mission of designing, deploying, and expanding a national cyberinfrastructure for life sciences research and training scientists in its use. The project defines cyberinfrastructure as a combination of hardware for computing and storage, software and tools, user codes, data, and people who provide training and support for a user base with varying skill levels.

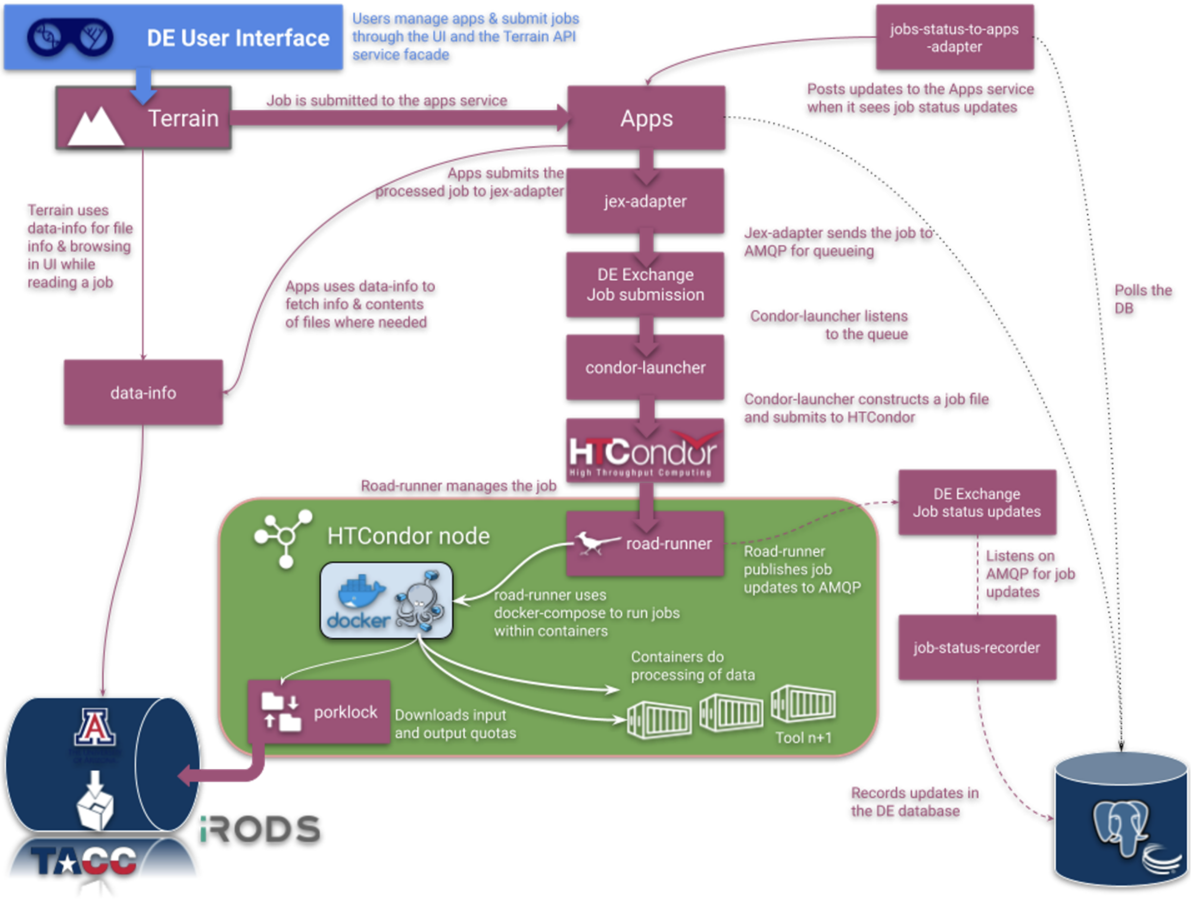
The core architecture of CyVerse is domain-agnostic and can be deployed on any platform, public or commercial, and can "connect" and leverage HPC private on-prem resources. At a high level, there is a layer of end-users "products" and a layer of services backed by a layer of hardware and cloud infrastructure, all integrated to give users and system operators a wide range of capabilities and options. The MDPS architecture is shown below:



The platform has an interactive development and "discovery environment" , which is deployed using K8s (Kubernetes) and the Cacao (https://gitlab.com/cyverse/cacao) system that "eliminate[s] the complexity of using multiple clouds," and enables researchers and educators to effortlessly manage, scale, and share their tools and workflows to any research-capable cloud using declarative templates. The IDE is shown below.



The IDE uses HTCondor (https://htcondor.org/) for HPC-oriented executable jobs (i.e., command-line executions) and workflows and integrates with the NSF's Extreme Science and Engineering Discovery Environment (XSEDE) for resource and on-prem usage tracking.


DE User Interface
Users manage apps & submit jobs through the UI and the Terrain API service facade
jobs-status-to-apps -adapter
Terrain
Job is submitted to the apps service
Apps
Posts updates to the Apps service when it sees job status updates
Apps submits the processed job to jex-adapter
jex-adapter
Terrain uses data-info for file info & browsing in UI while reading a job
Jex-adapter sends the job to AMQP for queueing
DE Exchange Job submission
Apps uses data-info to fetch info & contents of files where needed
Polls the DB
Condor-launcher listens to the queue
condor-launcher
data-info
Condor-launcher constructs a job file and submits to HTCondor
HTCondor High Throughput Computing
Road-runner manages the job
DE Exchange Job status updates
HTCondor node
road-runner
Road-runner publishes job updates to AMQP
Listens on AMQP for job updates
docker
road-runner uses docker-compose to run jobs within containers
job-status-recorder
Containers do processing of data
porklock
Downloads input and output quotas
Tool n+1
Records updates in the DE database
iRODS
TACC

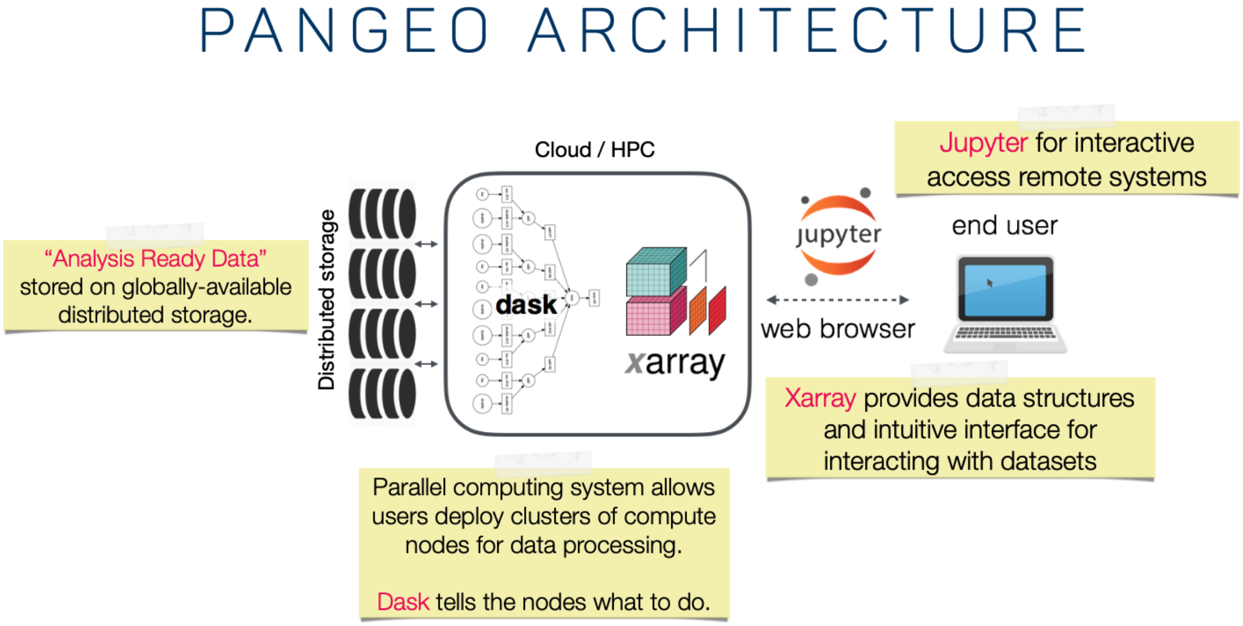
One core goal of CyVerse is to provide support and training. The need for workforce training was strongly highlighted in a survey of 704 NSF principal investigators (PI) under the Biological Sciences Directorate. Survey results showed that the most pressing needs are training in data integration, data management, and scaling analyses for HPC (Barone et al. 2017) to drive the analysis of relatively large biological datasets. NSF PI's noted that gaining access to a diverse set of computational and storage infrastructure is the lowest on their list of unmet needs.

In the spirit of Open Source Science, CyVerse holds workshops for a wide range of early career and general researchers who need cloud or HPC infrastructure for their work. In a workshop series on Foundational Open Science Skills, participants have learned to use version control, containers, and cloud – many for their first time. The success of the training is attributed to the accessibility of the end-users products within the discovery environment, or managed data science workbench, which is integrated with a variety of cloud and on-prem infrastructure and rich data stores.

Pangeo

Pangeo (https://pangeo.io) addresses some of the challenges with Big Data and science reproducibility, while closing technology gaps for the geoscience community through a unified, collaborative effort that combines open-source software, open data, and open infrastructure. The role of the Pangeo project is to coordinate scientists, software development, and computing infrastructure. The project facilitates an agile software development model where scientific users can provide immediate feedback to developers of open-source software libraries to drive meaningful development that meets the needs of the geoscience community. The discourse discussion boards (https://discourse.pangeo.io/) facilitate community communication for development and support. The expected impact is to provide the community with an integrated ecosystem of open-source software tools and an open Big Data platform that can scale to match the expected data growth of the NASA EOSDIS archive, which is projected to approach 250 PB by 2025.

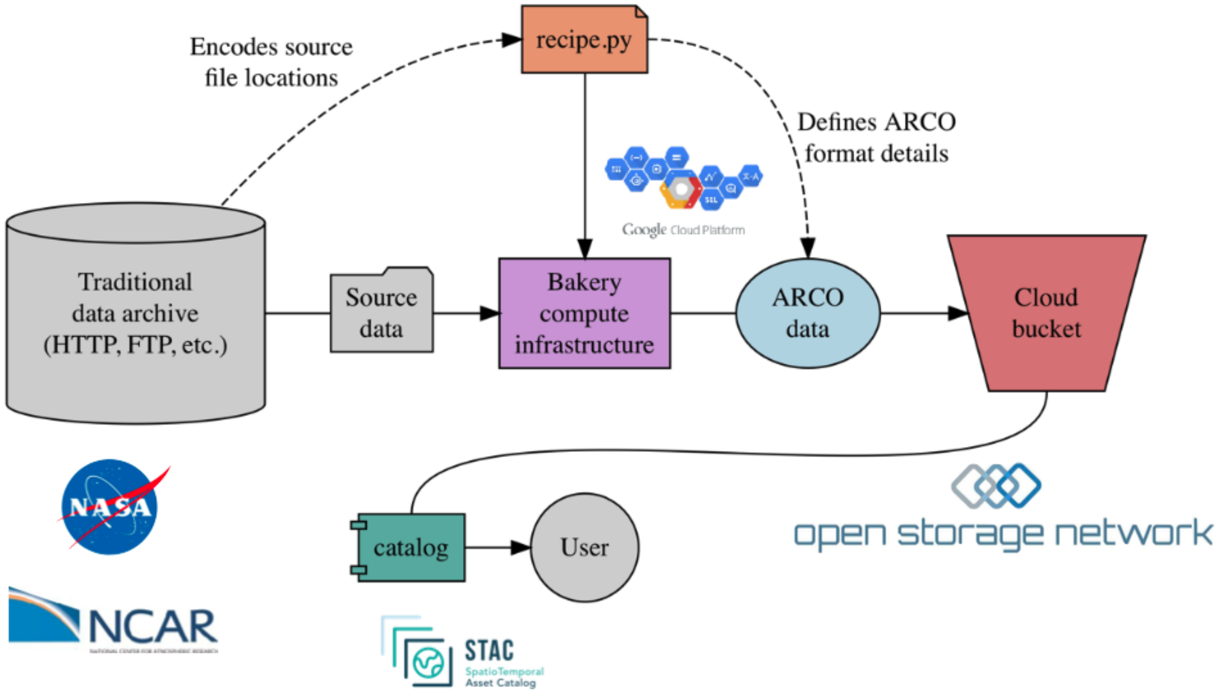
At a high level, the core components of the architecture for the open Big Data platform are Jupyter for interactive access to remote systems, Xarray to provide data structures for interacting with datasets, Dask for parallel computing, and deploying clusters of compute nodes for data processing. The open Big Data platform requires data to be ARD stored on globally-available distributed storage. The architecture is shown below.


PANGEO ARCHITECTURE
Cloud / HPC
Jupyter for interactive access remote systems
end user
jupyter
"Analysis Ready Data" stored on globally-available distributed storage.
Distributed storage
dask
xarray
web browser
Xarray provides data structures and intuitive interface for interacting with datasets
Parallel computing system allows users deploy clusters of compute nodes for data processing.
Dask tells the nodes what to do.

Many libraries and tools in the Pangeo ecosystem integrate with an instantiation of the platform including: Kerchunk (https://fsspec.github.io/kerchunk/), Pandas (https://pandas.pydata.org/), SciPy (https://scipy.org/), TensorFlow (https://www.tensorflow.org/), Pytorch (https://pytorch.org/), and cartopy (https://scitools.org.uk/cartopy/docs/latest/). A subproject called Pangeo-ML is working towards improving the interoperability of Pytroll and other ML

Libraries, which will enable seamless transitions between exploratory data analysis and machine learning applications.

Given that the open platform prefers ARD formatted data, a subproject, Pangeo Forge (https://pangeo-forge.org/), is now under development. Pangeo Forge is an open-source, community-driven platform for data Extraction, Transformation, and Loading (ETL) to make it easy for users to extract data from data repositories and deposit it in cloud object storage in analysis-ready, cloud-optimized (ARCO) format. Pangeo forge provides a high-level recipe framework alongside computing infrastructure to democratize ARCO data production. The ETL and STAC-based catalog generating process is shown in the figure below.

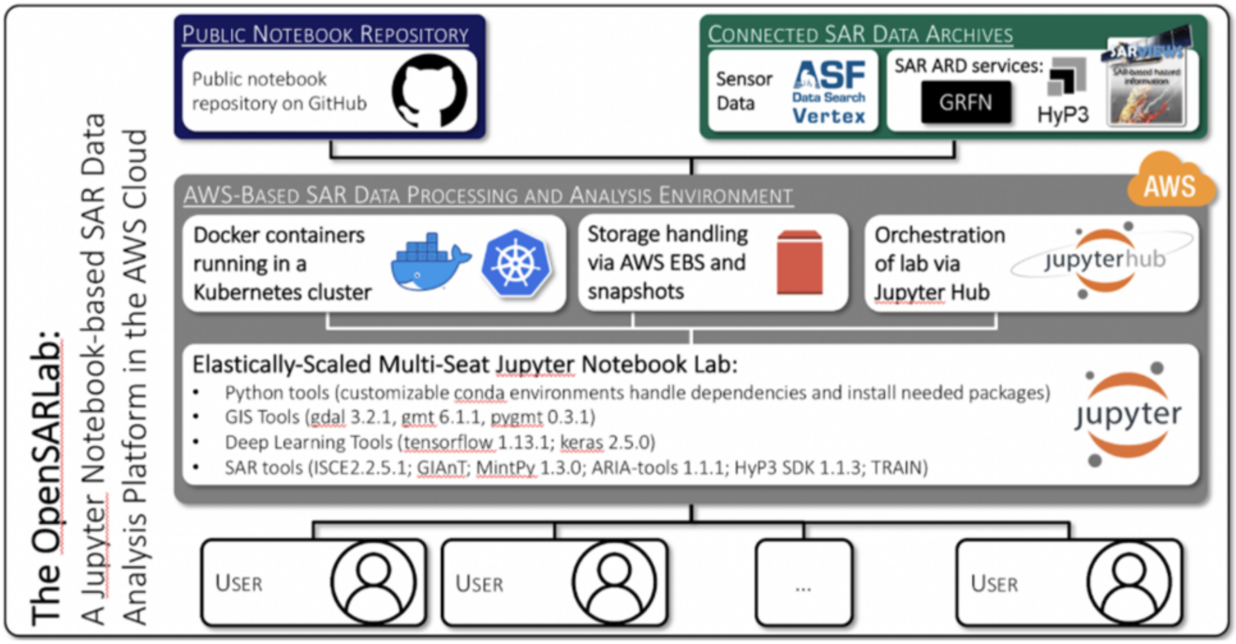

Encodes source file locations
recipe.py
Defines ARCO format details
Google Cloud Platform
Traditional data archive (HTTP, FTP, etc.)
Source data
Bakery compute infrastructure
ARCO data
Cloud bucket
NASA
NCAR
catalog
User
STAC
SpatioTemporal Asset Catalog
open storage network

Alaska SAR Facility

The Alaska SAR Facility (ASF) hosts the OpenSARlab, which is a service providing users persistent, cloud-based, customizable computing environments. Groups of scientists and students have access to identical environments, containing the same software, running on the same hardware. It operates in the cloud, which means anyone with a moderately reliable internet connection can access their development environment.

OpenSARlab sits alongside ASF's data archives in AWS, allowing for low latency transfer of large data products. OpenSARlab is a deployable service that creates an autoscaling Kubernetes cluster in Amazon AWS, running JupyterHub. Users have access to customizable environments running JupyterLab via authenticated accounts with persistent storage.

OpenSARLab has 3 components: Algorithm Development, HyP3 At-Scale processing, and Cumulus Data Ingest. The Algorithm Development is a Jupyter Notebook based SAR data analysis platform in AWS cloud. HyP3 At-Scale Processing is built on AWS Batch and is based on Dockers that provides fast and cost-effective processing. The Cumulus Data Ingest is a cloud-based data ingest, archive, distribution and management for Earth science data streams.


The OpenSARLab:
A Jupyter Notebook-based SAR Data Analysis Platform in the AWS Cloud
Public Notebook Repository
Public notebook repository on GitHub
Connected SAR Data Archives
Sensor Data
ASF Data Search Vertex
SAR ARD services:
GRFN
HyP3
SARVIEWS
AWS
AWS-Based SAR Data Processing and Analysis Environment
Docker containers running in a Kubernetes cluster
Storage handling via AWS EBS and snapshots
Orchestration of lab via Jupyter Hub
jupyterhub
Elastically-Scaled Multi-Seat Jupyter Notebook Lab:
• Python tools (customizable conda environments handle dependencies and install needed packages)
• GIS Tools (gdal 3.2.1, gmt 6.1.1, pygmt 0.3.1)
• Deep Learning Tools (tensorflow 1.13.1; keras 2.5.0)
• SAR tools (ISCE2.2.5.1; GIAnT; MintPy 1.3.0; ARIA-tools 1.1.1; HyP3 SDK 1.1.3; TRAIN)
jupyter
User
User
...
User

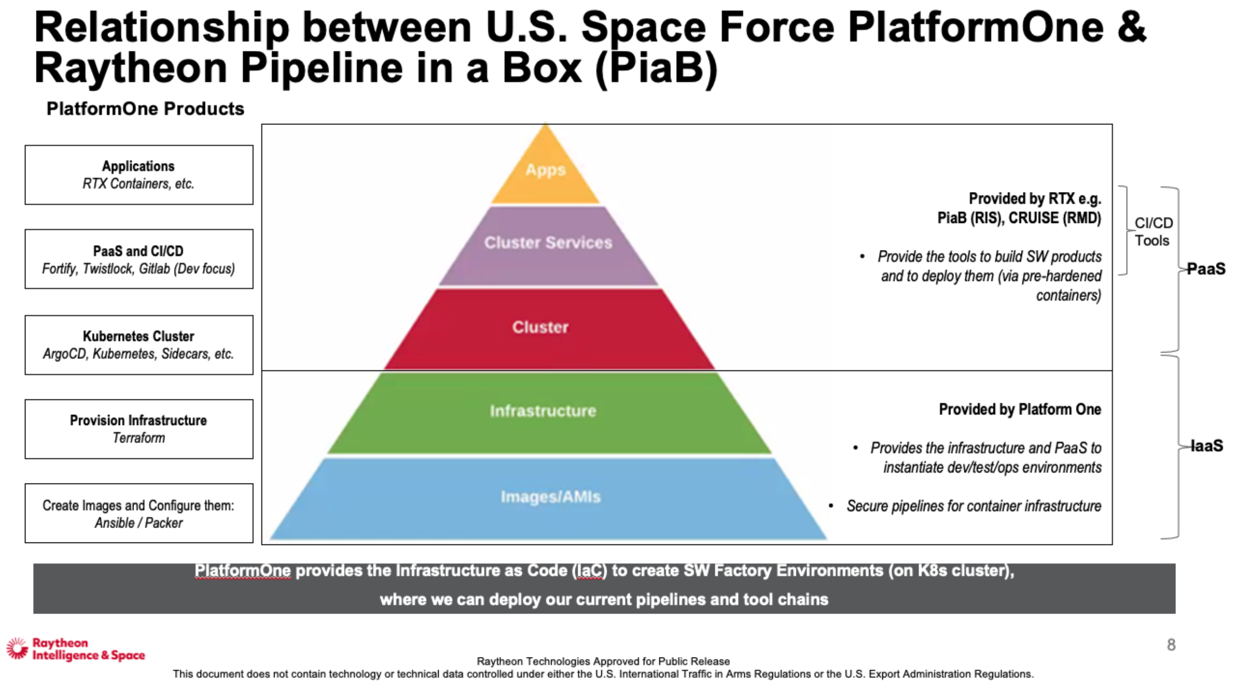
Raytheon - Capabilities for Big Data Processing Architecture

Raytheon has several capabilities and systems as existing building blocks that may be used to construct architectures that have the potential to accelerate Open Source Science innovation for ESO. The JPSS mission data processing is now wholly cloud-based and leverages some of the building blocks.

Raytheon has done work for NOAA's Earth Prediction Innovation Center (EPIC) Community Center (ECC), the Mission Data Processing Application Framework (MDPAF) for the USSF's Future Operationally Resilient Ground Evolution (FORGE) program, as well as other Internal Research and Development (IRAD) efforts at Raytheon. Additionally, the Pipeline in a Box (PiaB) system, which facilitates Agile and DevSecOps processes, was presented. The building blocks manage code repositories (algorithms, tools, documentation), compute resources (cloud-based and HPC compute), and data resources.

Open Source Science may benefit from the DevSecOps approach by providing automated testing that could significantly accelerate trying innovative solutions and offer greater repeatability for the broader community and the individual researcher. The use of structured processes like Agile and DevSecOps also encourages some rigor around the data used to drive the models, further enhancing repeatability and "apples to apples" comparison of innovative ideas.

Core technologies used to develop the various capabilities and how they relate to Raytheon's PlatformOne provide some architectural context. A conceptual framework is shown below:



Relationship between U.S. Space Force PlatformOne & Raytheon Pipeline in a Box (PiaB)

PlatformOne Products

Applications *RTX Containers, etc.*

PaaS and CI/CD *Fortify, Twistlock, Gitlab (Dev focus)*

Kubernetes Cluster *ArgoCD, Kubernetes, Sidecars, etc.*

Provision Infrastructure *Terraform*

Create Images and Configure them: *Ansible / Packer*

Provided by RTX e.g. PiaB (RIS), CRUISE (RMD)

- *Provide the tools to build SW products and to deploy them (via pre-hardened containers)*

Provided by Platform One

- *Provides the infrastructure and PaaS to instantiate dev/test/ops environments*
- *Secure pipelines for container infrastructure*

PlatformOne provides the Infrastructure as Code (IaC) to create SW Factory Environments (on K8s cluster), where we can deploy our current pipelines and tool chains

Raytheon Intelligence & Space

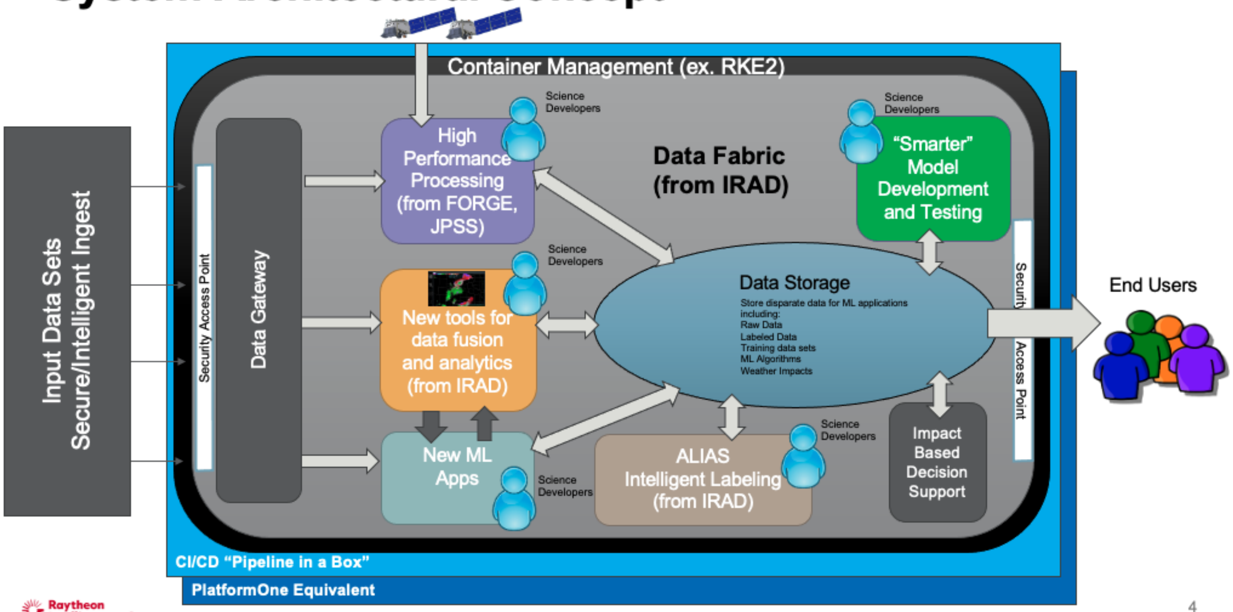
8

Raytheon Technologies Approved for Public Release
This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

A key component of the proposed architecture is the “Data Fabric,” an IRAD effort that aims to provide a single API that abstracts the details of the diverse storage mechanism that houses diverse types of data. The Data Fabric provides a simple, single point for data access for the system's users.

Additionally, the proposed architectural concept provides machine learning as a means of filtering data, supported by the Automated Labeling for Interactive Assisted Segmentation (ALIAS) system, which may reduce the burden on downstream legacy applications performing data assimilation. The ALIAS intelligent labeling "building block" from an ongoing IRAD effort addresses one of the most significant challenges for weather modeling systems.

System Architectural Concept


Container Management (ex. RKE2)
Input Data Sets
Secure/Intelligent Ingest
Security Access Point
Data Gateway
High Performance Processing (from FORGE, JPSS)
Science Developers
Data Fabric (from IRAD)
Science Developers
“Smarter” Model Development and Testing
Science Developers
New tools for data fusion and analytics (from IRAD)
Data Storage
Store disparate data for ML applications including:
Raw Data
Labeled Data
Training data sets
ML Algorithms
Weather Impacts
Security Access Point
End Users
New ML Apps
Science Developers
ALIAS Intelligent Labeling (from IRAD)
Science Developers
Impact Based Decision Support
CI/CD “Pipeline in a Box”
PlatformOne Equivalent

Raytheon Intelligence & Space

4

Raytheon Technologies Approved for Public Release
This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

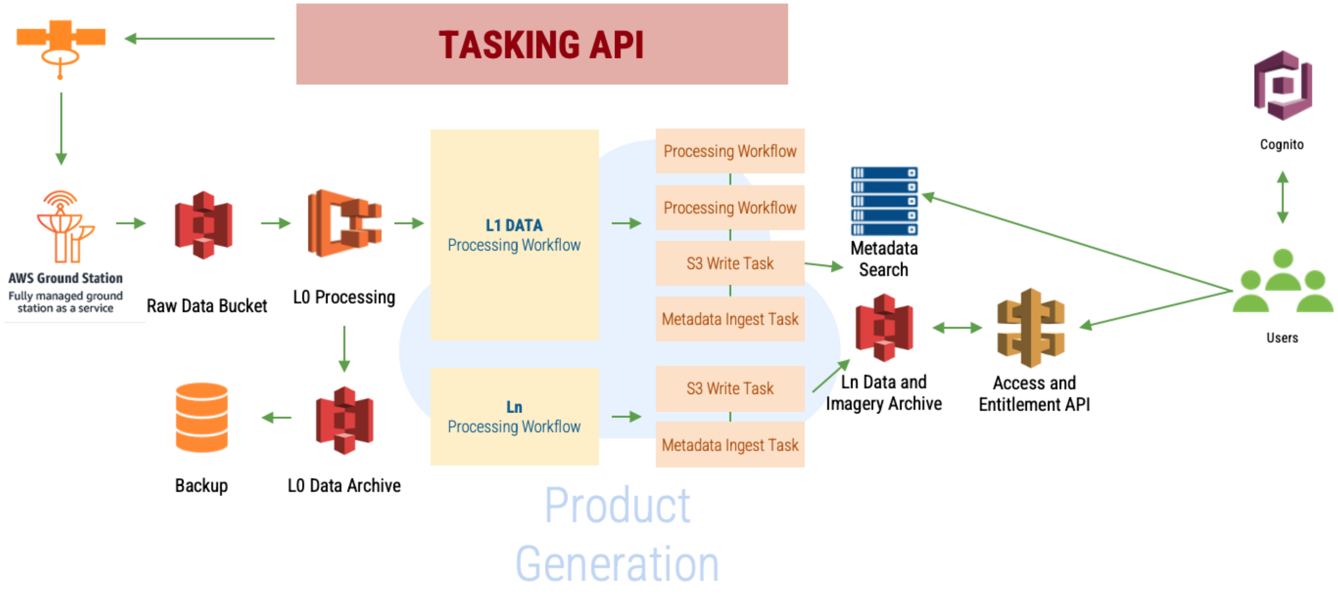
Element 84 - Perspective on Open Source processing systems

FilmDrop, an Element 84 product offering, is an open-source geospatial processing stack and data lake management solution. The system supports automated data ingest, archive, management, discovery, access, and processing pipelines for generating custom data products. Front-ends, such as dashboards, visualization tools, and Jupyter notebooks, can interface with data lakes and metadata repositories. The solution is easy to use, easy to deploy, and interoperable with many open-source software tools, systems, and APIs. The project is a spinoff of a NASA ACCESS award.

FilmDrop processing pipelines can incorporate private customer data, open and commercial data, and derived analysis products, which integrate into the data lake accessible via APIs and SpatioTemporal Asset Catalogs (STAC) metadata. STAC is a specification to describe geospatial data for search, discovery, and use in the cloud. It is JSON-based, crawlable (indexable), and has standardized metadata fields in records and record catalogs. A standard API for querying STAC, both server-based and file-based, can be used; an OGC API Features extension can also be used. Every data granule within FilmDrop has an associated STAC record. When appropriate, a processing field is added to a record, identifying the processor applied to a data granule, which provides the system's provenance-like capability.

The system aims to be cloud-native as much as possible and delegates infrastructure management and scaling to the CSP (AWS). Services such as serverless, AWS ground stations, managed object store (S3), and user management services (Cognito) were noted.

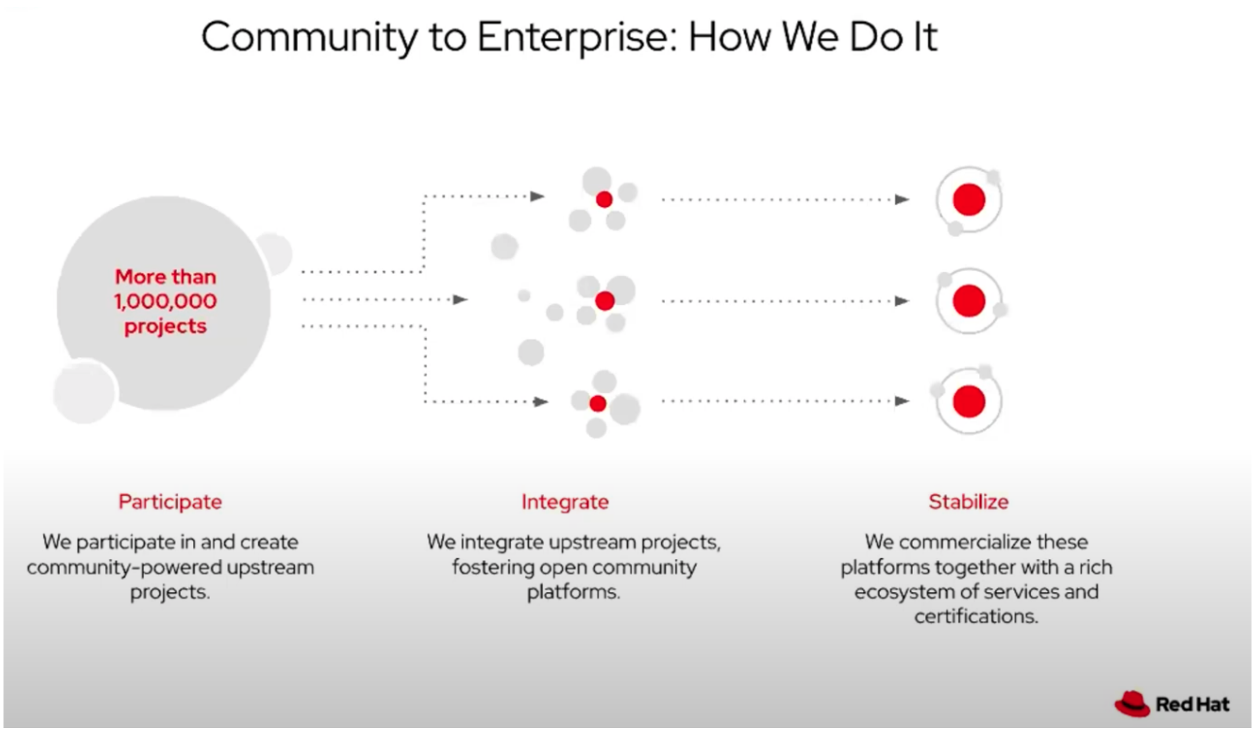
The FilmDrop architecture is shown below:



Red Hat - Open Source Methods and Philosophy

Red Hat, a stalwart in the open-source community, describes the process of how the company successfully provides enterprise open-source products and solutions. The company works with large, high-impact open-source community projects and productizes these into a fully supported, consumable, stable, and reliable, enterprise-grade solution for their customers. Red Hat does not currently outright own or develop technology or software "behind an internal wall," all products are derived from upstream projects, and all code produced by Red Hat engineers goes directly into the upstream community projects.

Participating, integrating, and stabilizing are the core three components of the Red Hat open-source philosophy and are how the company goes from "Community to Enterprise." Red Hat participates in thousands of community-powered upstream projects. A few notable projects in which Red Hat currently plays a significant role are Kubernetes (K8s) and OpenStack (https://www.openstack.org/), KVM (https://www.linux-kvm.org/page/Main_Page) and the Linux Kernel. Red Hat then integrates upstream projects and fosters open community platforms, which are then commercialized together with a rich ecosystem of services and industry certifications.



Community to Enterprise: How We Do It

Participate

We participate in and create community-powered upstream projects.

Integrate

We integrate upstream projects, fostering open community platforms.

Stabilize

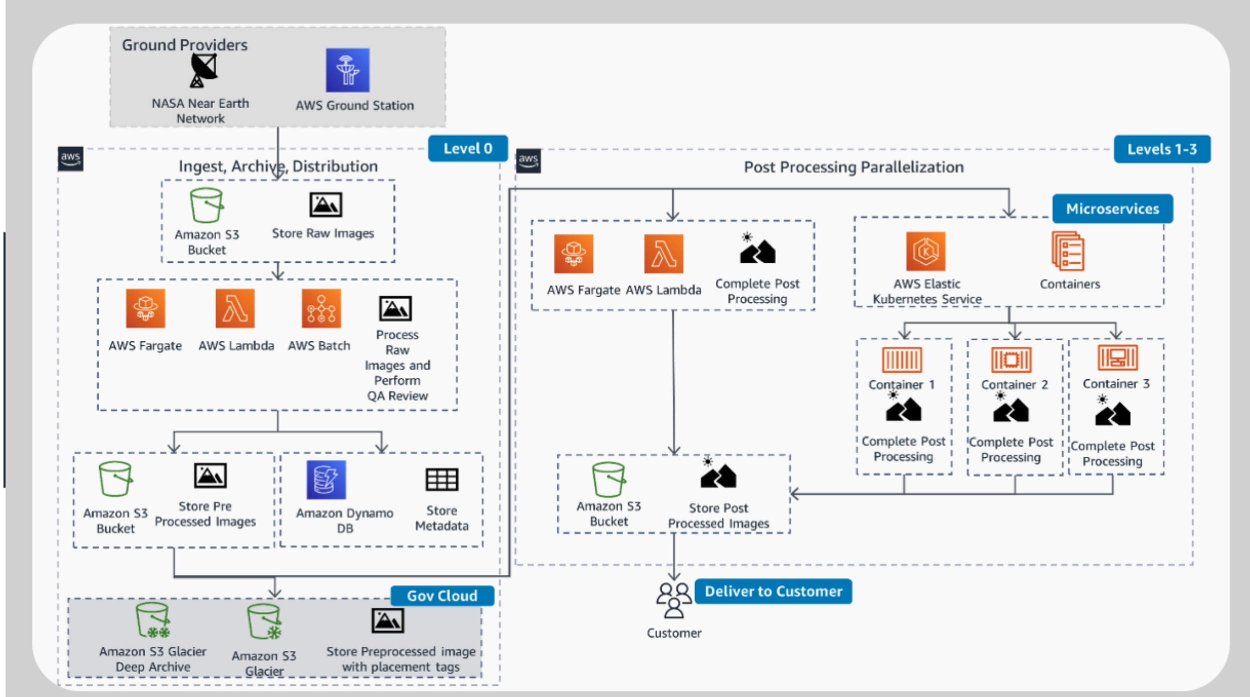
We commercialize these platforms together with a rich ecosystem of services and certifications.

The presenter suggested running the Red Hat development model (shown above) in reverse. Attempt to identify the most important and impactful technologies and build self-sustaining communities around those technologies. Strong governance, procedures, and stewardship are needed with the Open Source Science Initiative . Always be mindful of the open-source communities' development environments, tools, and needs, and consider fostering modern application development practices to achieve success.

Amazon Web Services - A Perspective on NASA Open Source Science ESO MDPS

Amazon Web Services (AWS) compute, storage, and services infrastructure can enable open science, reduce data processing time, and enhance time to mission science by bringing cloud capabilities of analysis and computing closer to the data sources thereby improving efficiencies to accelerate open source science.

AWS has a catalog of services, several worth mentioning in the ESO mission data processing context. The AWS ground station service provides a global downlink capability, integrates with the NASA Near Earth Network, and can rapidly facilitate data movement to the object store service, S3. With data downlinked and staged on S3, many AWS services, users and systems can consume the data from anywhere within the global infrastructure or externally, as permissible by security and governance policies. Services such as container management and orchestration, database services for cataloging metadata, and serverless computing capabilities are conveniently deployable as part of an MDPS or general users on the cloud interested in leveraging data assets. AWS enables architectures to take advantage of efficient data flows to relevant services and data locality. An example using some of these services is shown below:



Example of NASA JPL imagery architecture showing real-time data streaming and processing of JPL imagery.

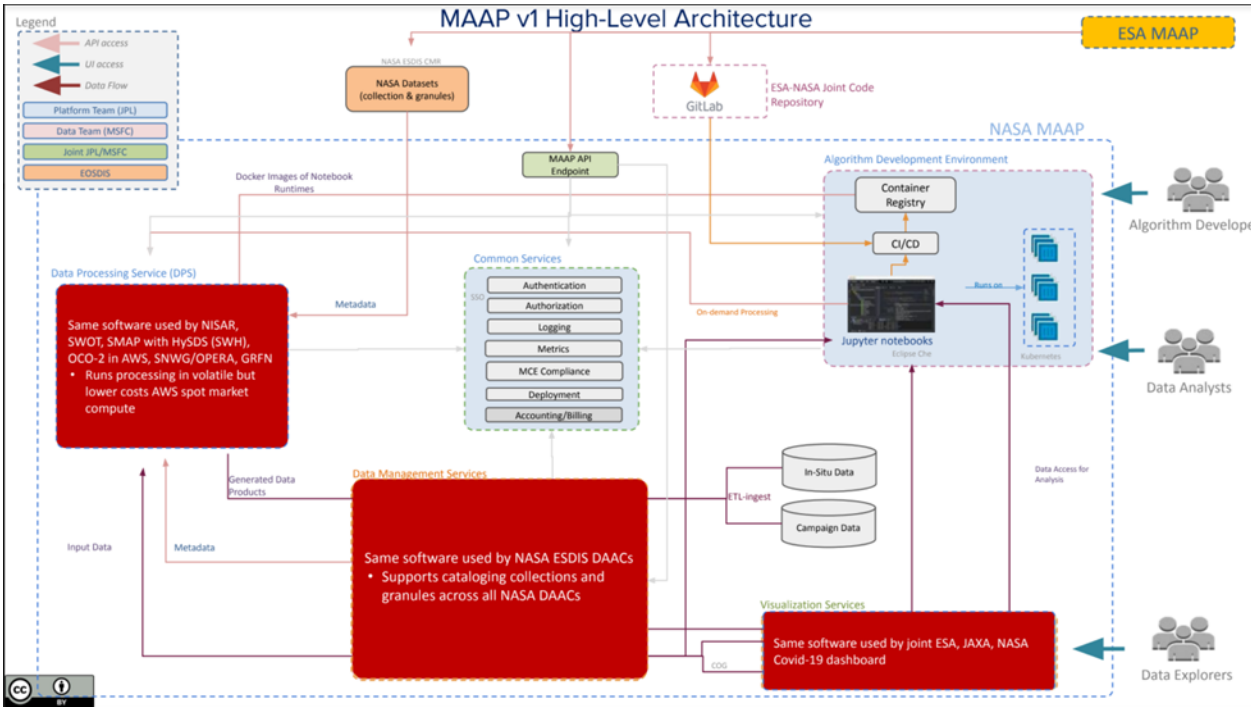
These services have also been used in a deployment of the open-source Advanced Multi-Mission Operations System (AMMOS) (https://ammos.nasa.gov/) for managing smallsat missions in AWS was shown as an example of using the available infrastructure resources, advanced open-source mission operations software, and community collaboration to enhance "time to science."

Multi-Mission Algorithm and Analysis Platform (MAAP)

MAAP is a cloud-based framework that provides a collaborative work environment that focuses on large satellite data stores along with code and runtime images for NASA and ESA scientists. Some of the services MAAP provides to their customers are algorithm development, metadata management, dashboard, data processing service, and code sharing.

The algorithm development environment uses Eclipse Che which is a cloud-based Kubernetes native Interactive Development Environment (IDE) manager. This service can run JupyterLab, RStudio or other web-based IDE's for traditional programming. MAAP utilizes NASA ESDIS services like Cumulus, CMR, and Earthdata Search (See NASA System Interface) as well as the European Space Agency (ESA) Metadata Management Tool (MMT) for data cataloging. The dashboard was developed for the COVID-19 partnership between NASA, ESA and JAXA that uses GitHub for automated deployments and dataset configurations. The Data Processing Service (DPS) executes prebuilt containers with algorithms developed in the ADE. DPS is currently leveraging Amazon Web Services (AWS) auto-scaling capabilities but is looking at other venues for large scale processing.

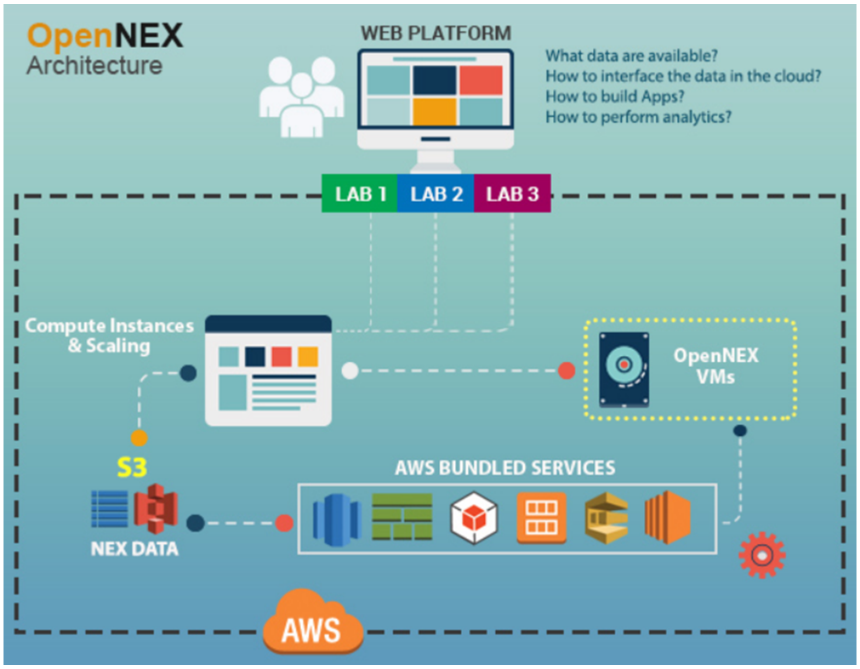
The MAAP architecture is shown below:



OpenNEX

The NASA Earth Exchange (NEX) is a program to enable scientific collaboration with bit data next to compute thereby reducing the community's need to move large datasets and facilitate sharing of redundant code and workflows. Some of the challenges NEX has overcome are finding, ordering, waiting, downloading, pre-processing, and performing large scale computing with big data. Some of the new challenges are security constraints and onboarding logistics. OpenNEX is the public cloud version of NEX.

The OpenNEX architecture is shown below:



NEX and OpenNEX have become platforms supporting scientific collaboration, knowledge sharing and research for the entire Earth science community. To date, a number of custom tools and capabilities have been integrated into the platforms. However, such integration has to undergo a case-by-case manual process that lacks scalability. This timely project builds an App Store onto OpenNEX as a building block. Climate data analytics tools/programs can be easily uploaded, shared, organized, searched, and recommended like photos and videos on YouTube. The foundation of the App Store is a provenance server, which not only records metadata but also execution history of climate data analytics apps including: input data and parameters, output data and products, who runs the app for which purpose, and how apps may be chained into workflows. Researchers can thus understand, reproduce, and repurpose existing apps and workflows. Machine learning approaches are used on provenance metadata to provide recommendations on as-you-go services (e.g., apps and workflows) for Earth scientists. A browser-based workflow tool is also provided for researchers to explore the provenance server and design value-added workflows. Scalability, sustainability, extensibility, usability, adaptability, security and privacy are considered in the App Store.

Unity (Science Data System as a Service)

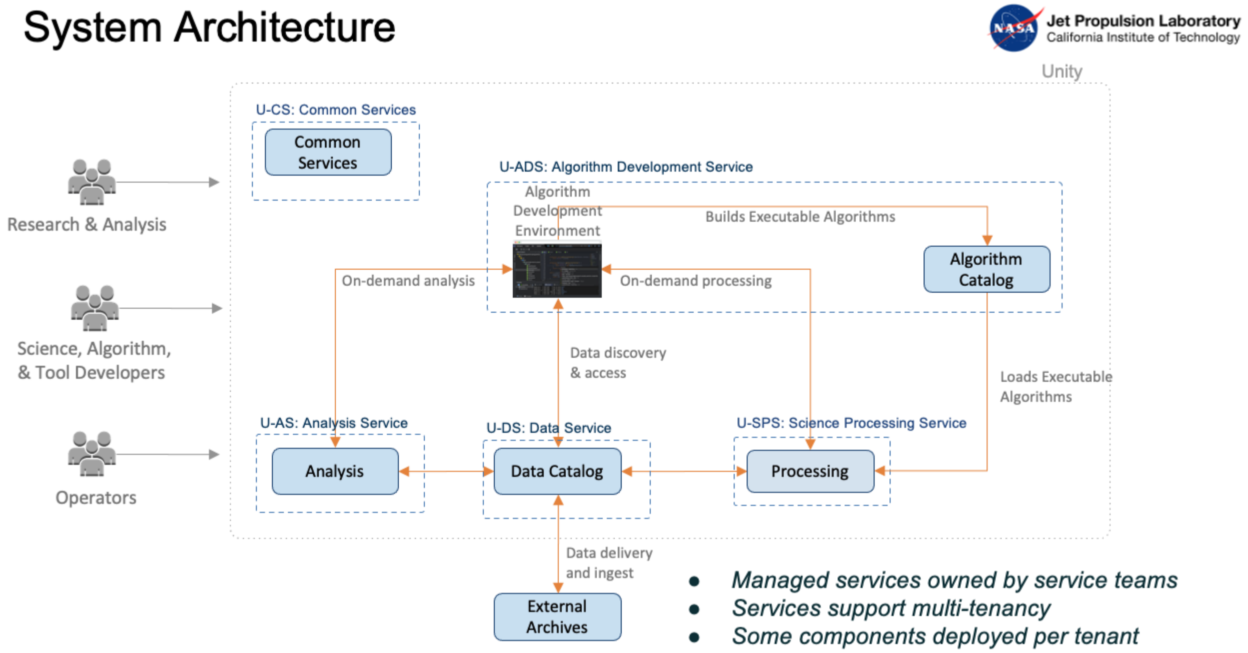
The Unity platform is a next-generation Science Data System (SDS) that is service-based and focused on mission science data processing. Platform users can develop, test, and validate product generation algorithms in an operations-like environment, improving the continuous integration and development cycle thereby facilitating technology infusion in the algorithm development and data analysis space.

The platform provides a multi-tenancy system supporting multiple customer projects at any given time. The customer-focused SDS simplifies the onboarding process and reduces project SDS deployment times and costs through reuse and strong collaboration, all organized around dedicated service-oriented business teams. The platform capabilities provide a suite of managed services to streamline the processing, storing, and managing of science data products. Each service has a group that owns it, and that group is responsible for developing, maintaining, maturing, and evolving that service to meet customer needs.

Unity implements standards-based, interoperable services and algorithms to enhance cross-project collaboration and promote work portability and reuse. This approach allows loosely coupled services to work together to achieve a customer's goal. Unity has identified Open Geospatial Consortium (OGC) (https://www.ogc.org/) standards and OGC application packages' best practices (https://docs.ogc.org/bp/20-089r1.html), WPS-T (https://docs.ogc.org/per/18-036r1.html), a transactional web processing service (WPS), and other promising standards as a way to improve interoperability for data processing and related components, as well as create execution of services effectively.

The system architecture and its components are shown below.

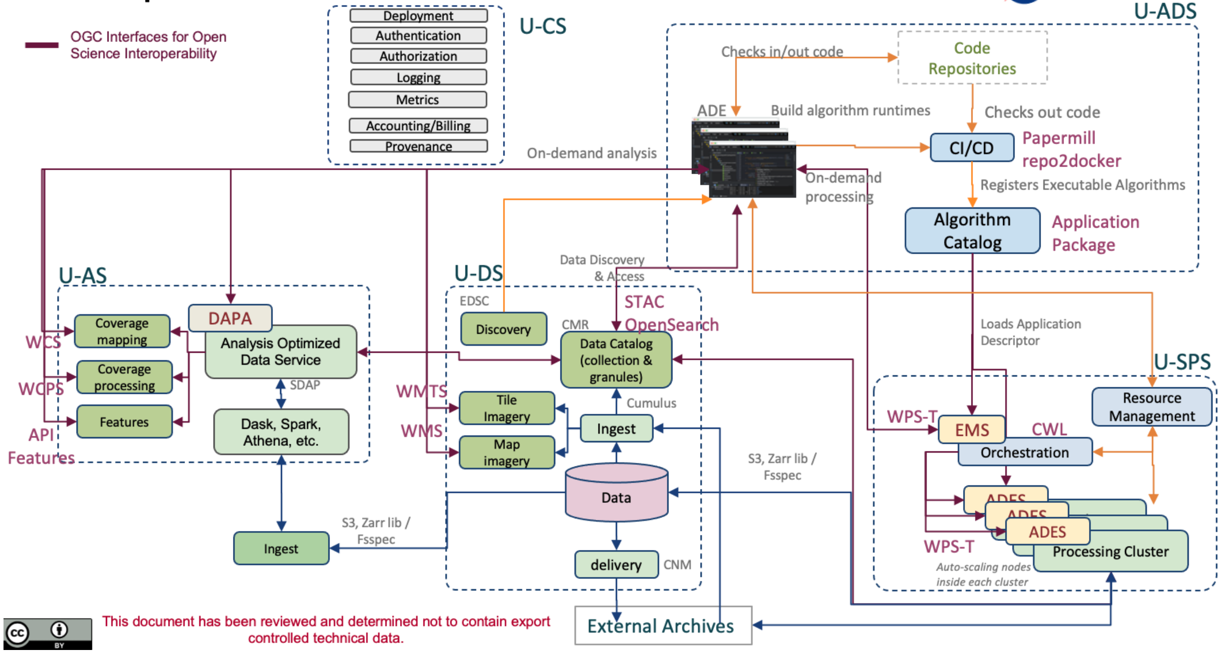
System Architecture



- *Managed services owned by service teams*
- *Services support multi-tenancy*
- *Some components deployed per tenant*

Component and Infrastructure View

Jet Propulsion Laboratory
California Institute of Technology

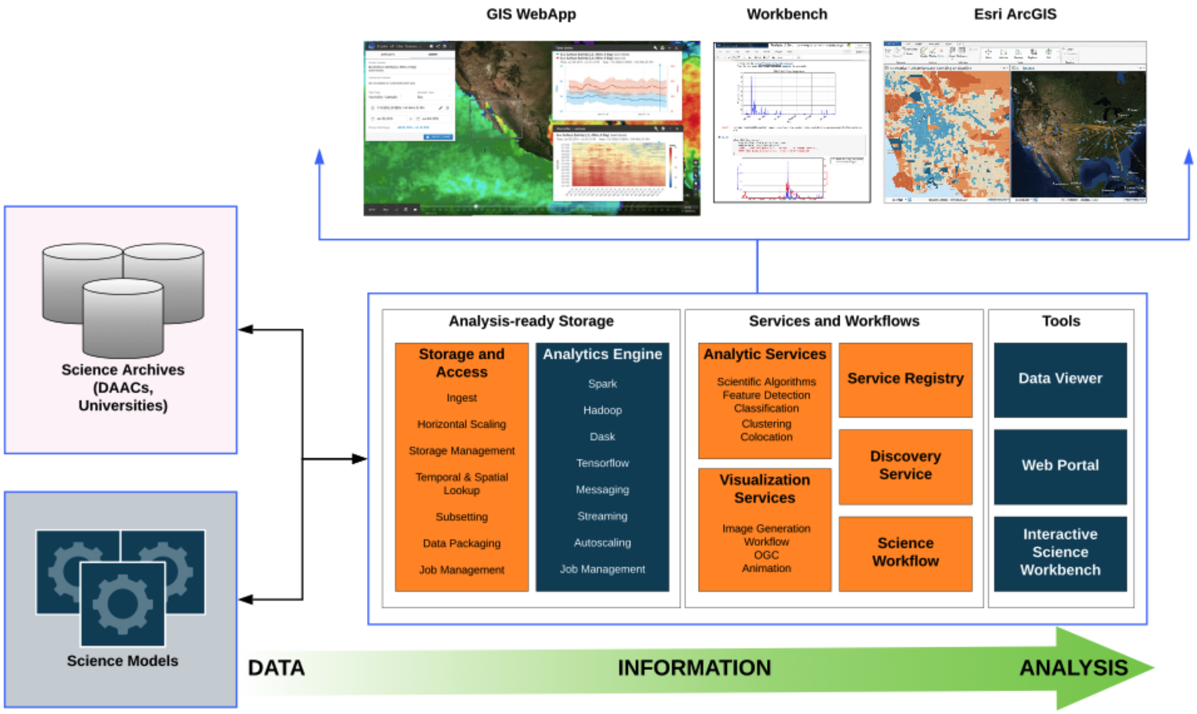


This document has been reviewed and determined not to contain export controlled technical data.

5

Science Data Analytics Platform (SDAP) - Apache Science Data Analytics Platform Perspective

The Apache Science Data Analytics Platform (SDAP) is a professional open-source project that builds and maintains an Analytics Collaborative Framework (ACF). The platform provides a data and tool-rich environment for conducting science investigations that can be tailored to individual study areas, such as the physical ocean, sea level, and air quality. APIs for data access and ingest and tooling are available and supported by compute resources for tasks such as harmonizing data. SDAP provides infrastructure and data management capabilities so researchers can focus on performing scientific investigations and collaborate with peers.



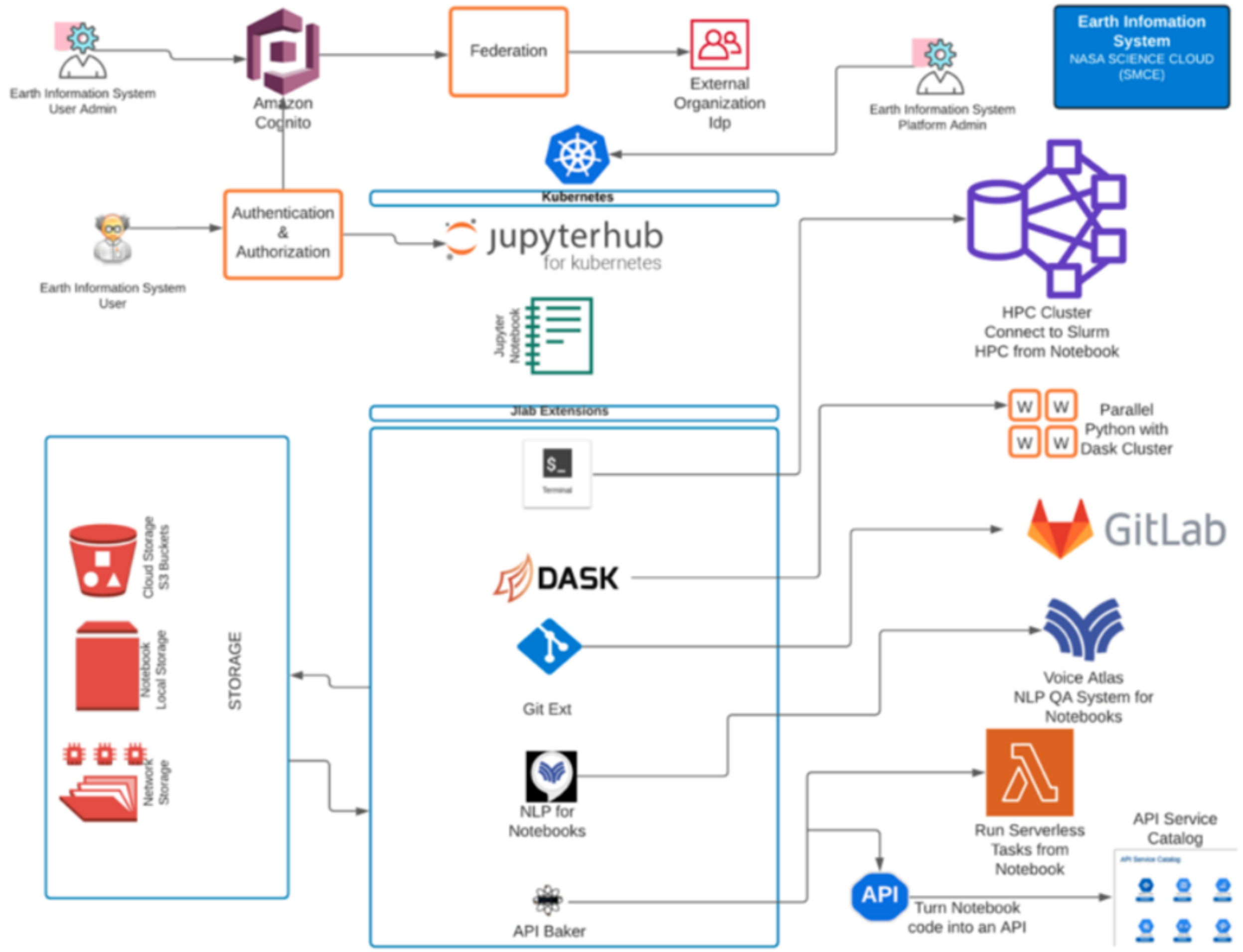
The system streamlines deployments using several technologies. Terraform (https://www.terraform.io/) is used for provisioning infrastructure. Kubernetes, YARN (https://hadoop.apache.org/docs/stable/hadoop-yarn/hadoop-yarn-site/YARN.html) or MESOS (https://mesos.apache.org/) is used for scaling and or managing containerized applications. For Software packaging and management, container-based technologies and Helm (https://helm.sh/) charts are used. SDAP can be deployed on-premise and in the cloud, and the project team strives to build a multi-cloud, multi-cluster, multi-data-center, and multi-agency system. By streaming deployments on various infrastructures, it may be more straightforward for organizations to adopt SDAP, and the project can achieve better system uptake.

The system uses data stores such as Apache Cassandra, ScyllaDB, and AWS S3; for spatial indexing, the system uses Apache Solr or ElasticSearch. Analytic Engines, such as Apache Spark Cluster (https://spark.apache.org/) or Amazon Elastic MapReduce (EMR), enable large-scale data analysis.

NASA Earth Information System (EIS)

The NASA Earth Information System (EIS) is a program development environment enabling access to the NASA-owned AWS organization named Science Managed Cloud Environment (SMCE). The program has tackled complex Earth science decadal survey (NRC, 2019) questions by synthesizing state-of-the-art models and observations from NASA and its partners. EIS facilitates interdisciplinary scientific collaboration and stakeholder engagement leveraging open-source tools and emerging computing capabilities. It helps translate scientific results into actionable information for a wide range of users and stakeholders. SMCE is a low-security AWS environment that aggressively uses AWS EC2 spot instances for cost effective data processing. GitLab is utilized for software version control and collaboration for engineers and scientists. SMCE also has a team of systems administrators that monitors costs using AWS Cloudwatch.

The EIS architecture is shown below:



Earth Information System User Admin
Amazon Cognito
Federation
External Organization Idp
Earth Information System Platform Admin
Earth Infomation System
NASA SCIENCE CLOUD (SMCE)
Kubernetes
Authentication & Authorization
Earth Information System User
jupyterhub
for kubernetes
Jupyter Notebook
HPC Cluster Connect to Slurm HPC from Notebook
Jlab Extensions
W W W W
Parallel Python with Dask Cluster
Terminal
GitLab
DASK
Git Ext
Voice Atlas NLP QA System for Notebooks
Cloud Storage S3 Buckets
Notebook Local Storage
Network Storage
STORAGE
NLP for Notebooks
Run Serverless Tasks from Notebook
API Service Catalog
API Baker
API
Turn Notebook code into an API

7.3 Key Points from Breakout Session Discussions

7.3.1 System development approaches & challenges

This breakout session occurred on Days 1 and 2 of the workshop with variable participation each day and covered topics such as: development challenges, teams, reuse, open source software development, I&T phasing, cybersecurity, cost model, etc. Important takeaways from this discussion included the need to separate resources used in the cloud. It was noted that AWS Pcluster can help allocate resources; however tracking resources and accounting for use could be a challenge. Another challenge is how security is handled, which differs from traditional on-premise systems that have extra security measures through an on-boarding process.

When designing an architecture it is important to consider variations on cloud architectures; for example, use of shared resources and different cost models for each resource. It is important that shared resources across systems operate within the same cloud facility (e.g., AWS region).

Important to considering different architectures, and shared services is a need for a common terminology of subsystems and components as the building blocks. This would make reuse easier.

Consideration of these aspects is important to enable broader use of NASA data as users have very diverse needs, many beyond the ability of a flight project to provide. Reusability and broader access enables response to the diversity of those needs.

7.3.2 System operations approaches & challenges

This breakout session occurred on Days 1 and 2 of the workshop with variable participation each day and covered topics such as: architecture deployments, concept of operations, roles, number of customers and platforms, shared services, integration and testing, cybersecurity, cost models, etc. One of the biggest challenges to system operations is providing access for external partners. There are software patches that can help, but these have impacts on security, cost, and schedule. Cybersecurity puts a burden on operations teams, which bring unfunded mandates to the system. The NOAA NCCF system is a promising model for centralizing and achieving efficiency, but it also raises issues such as cost sharing, management, and technical coupling. It may be most useful to ask what features users want from an operational system, which can be typically characterized as lower costs and latency as well as analysis ready data (ARD), but this is a loaded term that is specific to each application. Other challenges include finding and maintaining a workforce with the right skills to maintain the system.

7.3.3 Open Sourced-Science approaches & challenges

This breakout session occurred on Days 1 and 2 of the workshop with variable participation each day and covered topics such as: community code contributions, sharing code, accelerated analytics, improving participation from underrepresented communities, on-demand and community-generated product generation, multi-source data, cross-cloud, cybersecurity limitations, etc. With respect to open-sourced science and an MDPS, there is a need to define criteria or a process for community contributions. Specifically, how are algorithms and products (including at different maturity levels) reviewed to determine which products should be certified by NASA. An important component of what NASA data products provide is trustworthiness of algorithms and quality assurance, both of which have become expectations from users. A process is needed for how the user community can contribute vetted data back into the system. Also needed are metadata standards and documentation for provenance tracking and how to extend this to include community contributions. We don't currently have a model for storing data that has already been vetted (e.g., ISO 16363 freely available through the Consultative Community for Science Data Systems - CCSDS) that tracks all changes and revisions since inception. There was discussion of the use of an open-source contributor model for design of science teams, rather than the traditional NASA awarded science teams. To truly enable community collaboration and participation, the highest value algorithms need to be owned by the community, who can contribute and maintain the algorithms through time. This requires thinking about how to make proposal contributions easier and provide seed/maintenance funding for community contributions. This would be necessary to successfully change the culture around open science.

To enable open science, an MDPS needs mechanisms for allocating costing (e.g. include sandbox costs and hidden overhead costs like egress and operations and maintenance), reconsidering internal policies of vetting algorithms, and recognizing contributions (e.g., mechanisms for capturing contributions and crediting contributors). To enable better mechanisms for cost allocation, there were discussions recommending a cost model study to determine breakdown of costs to contributors and NASA, not just cloud resources, but costs for getting and maintaining existing quality standards for contributions from science teams. While infrastructure exists for vetting and peer reviewing algorithms, it is unclear how this would be implemented within an MDPS architecture. Finally, contributors need recognition for the work that they do, and this requires considerations of mechanisms that both enable tracking and ethics to user-ownership (e.g., CARE data principles; Carroll et al., 2020).

7.3.4 Open-Source Software approaches & challenges

This breakout session occurred on Days 1 and 2 of the workshop covered several topics such as: incentives for open source, legacy systems, proprietary code/systems, improving participation from underrepresented communities; incentivizing contributions, contribution "verification" quality, intellectual property, licensing, true costs, standards (coding, etc..), and long-term support. There was variable participation for each of the two breakout sessions.

Open-Source Science can leverage the lessons learned from the open-source software community, by learning from and contributing to existing tools and efforts. Specifically, open-source software has provided many lessons learned with respect to licensing issues, using repositories, classifying authoritative software, costing to support modifications, making software interoperable for different platform compatibility, and lacking clear communication on the processes for community contributions. Open-source software is not just about making code publicly available and following FAIR data principles (Wilkinson et al., 2016), it is about communication and creating a culture. Stakeholder communication and needs capture will help to turn a community of users into contributors.

There is a lack of clear guidance on policies and procedures to reach OSS goals as many NASA policies are directly oppositional, contradictory. Most importantly, there needs to be documentation and buy-in at all levels on policies and procedures from top-level HQ all the way down to the developer. This would help with addressing valid issues like security by developing a security approach that is both compliant and implementable. At present, the NASA process is inconsistent with NASA policy - it is unnecessarily complicated and ill-defined. This makes it challenging to meet OSS goals without significant changes in processes. A working group should look at Agency policies and processes and implementation to make recommendations.

7.3.5 Data Analysis needs from an MDPS

This breakout session occurred on Days 1 and 2 of the workshop with variable participation each day and covered topics such as: science teams algorithm development, product generation, quality assurance, cal/val, supporting science and applications from end products from the MDPS, improve participation from underrepresented communities, user communities/stakeholders, on-demand product generation, community-generated higher-level products, etc.

There was discussion on the need and importance of analysis-ready data (ARD) recognizing a potential trade-off between mission products, on-demand products and ARD products, especially with cost-caps in place for any single mission. Important to ARDs is the reproducibility and openness of their production and the use of shared resources/reusability of the tools in data product generation and analysis, that can support ARD production for different measurements from disparate missions hosted by different centers.

To do this there needs to be a feedback mechanism between closed development environments for low level products and open environments for high-level products in an MDPS. At present, we are missing the feedback loop between the two. To accommodate this, we need more engagement with the science team and the broader community early in the mission as well as a change in the data management approach that goes beyond a build-test-operate paradigm to a more iterative paradigm that integrates improvements during operations using a more agile process for contributing code. This will require plans for improving algorithms in the operations phase that is agreed upon during the development phase of the mission.

Another dimension discussed was how different centers running multiple missions have separate software bases and stop improving after commissioning. What’s needed is convergence on reusable components that can be used in more custom-built solutions and can be improved through open source.

HECC can play a role in supporting these efforts especially as it can be difficult to move users into the cloud, which requires training.

7.3.6 MDPS Architectures Now & in the Future

This breakout session occurred on Days 1 and 2 of the workshop with variable participation each day and covered topics such as: interfacing with the community from the cloud (DAACs, analysis platforms, etc), cloud economics, Managed Services,Standards based interoperability, quality control, licensing, etc.

There is a need to identify the “common denominator” of an MDPS that addresses the fundamental services in the architecture. This would help delineate some of the overlaps in architectures for “before the archive” (MDPS) and “after the archive” (DAAC) analysis. The key is to delineate the role of analytics in MDPS as an iterative assessment and refinement of data products (e.g. GEDI SOC) and for the use and re-use of MDPS services for on-demand processing. This opens the conversation to “science-as-a-service” and how an MDPS architecture may support agile usage of services that facilitates the building of higher-level science data products. Important to this consideration is what is the “color of money” funding and how that falls within scope for each budget. This requires well-defined interfaces both for users (algorithm developers and scientists as well as MDPS developers. To define these interfaces, it’s necessary to work with the science community rather than enforcing approaches familiar to the developer community.

Within this common denominator MDPS there needs to be consideration of common services versus centralized services. For example, cumulus is a common capability that is deployed by different projects, while CMR is a centralized service for which different projects contribute.

There was also recognition of multi-cloud equivalence of services and implications for vendor lock-in. One option is to build interoperability into the MDPS design thinking about services that have equivalence across vendors and use of containers for algorithms.

Interoperability across cloud vendors is not as straightforward between HPC and cloud. Specifically, a “lift and shift” approach to the cloud does not work as well and there is a need to embrace more cloud-native approaches (e.g., serverless, etc.). There is a need to apply a cloud-native development process, recognizing that there exists many reusable tools that are not AWS-centric (e.g. terraform vs cloudformation) and how these may port to HPC.

7.3.7 Defining an MDPS

On Day 3 of the workshop, a more organic approach was taken using mentiment to gauge topics that participants felt had not been adequately addressed using an open-ended discussion topics feature. As topics were posted by the participants, the SAWG invited people to speak up and elaborate on different topics. One topic that generated a lot of discussion was how to define an MDPS.

An MDPS processes data for product generation with mechanisms in place to determine if products are ephemeral or worth long-term archive. It is worth noting that product creation could be by the public or ST, and this is not necessarily predefined. Products can include: 1) standard NASA project products that are within scope for a project to generate; 2) on-demand products from NASA-approved algorithms and workflows; 3) on-demand products of customized variants of NASA approved algorithms (e.g., locally calibrated); and 4) non-NASA products (e.g., state level - as opposed to global, applications focused, etc.). There was recognition that there is non-trivial skill and resources to generate products and access by the broader community to an MDPS could really expand the use of the NASA data, but that it was also a non-trivial cost for user support should more users have access to MDPS services.

A DAAC provides long-term data archive with user services and support that ingest data from an MDPS and interface with analysis platforms. DAACs span missions focusing on a specific scientific area and have been in place for many years. They have evolved over the years as NASA evaluates the DAAC content ensuring continual evolution to meet changing needs through time. Data products move from an MDPS to a DAAC as per SPD-41 (NASA, 2021).

An Analysis Platform is used for scientific analysis of products generated and made available, not for generating products.

7.3.8 On-Prem versus Cloud

On Day 3 of the workshop, a more organic approach was taken using mentiment to gauge topics that participants felt had not been adequately addressed using an open-ended discussion topics feature. As topics were posted by the participants, the SAWG invited people to speak up and elaborate on different topics. One topic that generated a lot of discussion was the use of on-prem HPC vs cloud options.

There seems to be very strong and polarizing perspectives on cloud versus on-premise HPC. The case for use of on-premise HECC is related to costs as it costs less to use HPC and that NAS has shown a preliminary cost saving estimate when using HECC compared to AWS (Hood & Jin, n.d.). However, it was recognized that there is a need to conduct true Total Cost of Ownership (TCO) analysis both to NASA and to the user of the system that accounts for all costs including “hidden costs” such as egress from cloud, workforce, and system operations and maintenance. The case to use cloud was related to: 1) access in support of open science (i.e., no on-boarding); 2) access to the latest computing technologies (e.g., GPU, TPUs, etc.); and 3) the ability to bring your own resources to expand the resources any user may need independent of whether the funding comes from NASA.

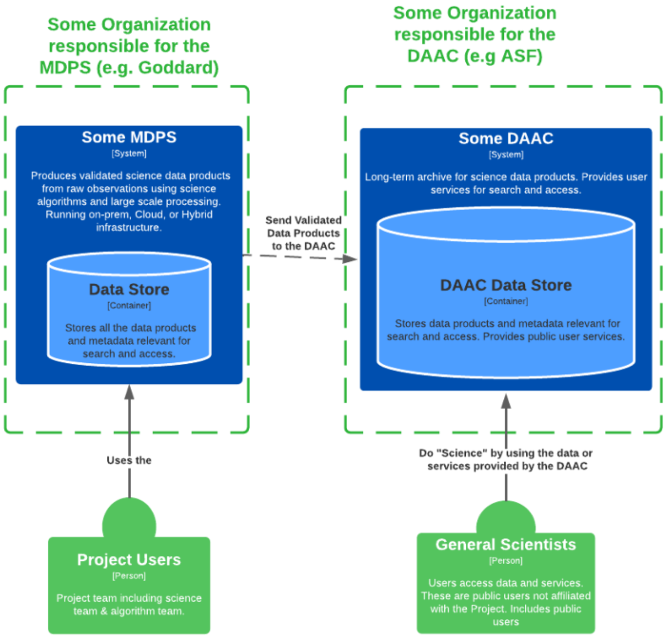
7.4 Synthesis of Findings to Define a MDPS

One of the key messages from this workshop was the need for a clear definition of a Mission Data Processing System (MDPS). To do this, the SAWG first synthesized the common architectures presented and then created a block definition diagram (BDD) that does not show how parts interact with each other, but rather clearly outlines the necessary parts for building an MDPS.

7.4.1 Synthesis of Common Architectures from Workshop 2

There were three common architectures across all of those presented: 1) Single instance with the MDPS sending data to the data archive center representing the heritage approach where an MDPS and DAACs are independent systems and may be distributed with defined data product delivery interfaces; 2) collocated MDPS and data archive; and 3) multi-mission MDPS with multi-tenant analysis environments. The majority of the system architectures presented shared a common set of functional capabilities and components for data management, scalable data processing, and algorithm development. There were notable themes of leveraging hybrid AWS cloud and HECC on-premise platforms. Differences in the platforms presented were mainly in the technology implementations and deployments. Some architectural approaches were not only moving to cloud, but also consolidating common components to a managed architecture to support a multi-tenancy approach to science data processing.

We present diagrams for each of these common architectures below.



Some Organization responsible for the MDPS (e.g. Goddard)

Some MDPS
[System]
Produces validated science data products from raw observations using science algorithms and large scale processing. Running on-prem, Cloud, or Hybrid infrastructure.

Data Store
[Container]
Stores all the data products and metadata relevant for search and access.

Send Validated Data Products to the DAAC

Some Organization responsible for the DAAC (e.g ASF)

Some DAAC
[System]
Long-term archive for science data products. Provides user services for search and access.

DAAC Data Store
[Container]
Stores data products and metadata relevant for search and access. Provides public user services.

Uses the

Project Users
[Person]
Project team including science team & algorithm team.

Do "Science" by using the data or services provided by the DAAC

General Scientists
[Person]
Users access data and services. These are public users not affiliated with the Project. Includes public users

Architecture 1 - Single Instance (This seems to be the way most systems are architected)

Each MDPS is a separate instance independent of other MDPS's the organization may implement

Each instance of an MDPS mostly uses common components and common open source technologies, but those are generally only shared within one ogranization (i.e. Goddard, JPL, NOA, USGS, DLR, etc. but not shared across organizations)

The MDPS maintains its own copy of all the Data, any additional test data is also stored at the MDPS and generally not transfered to the DAAC.

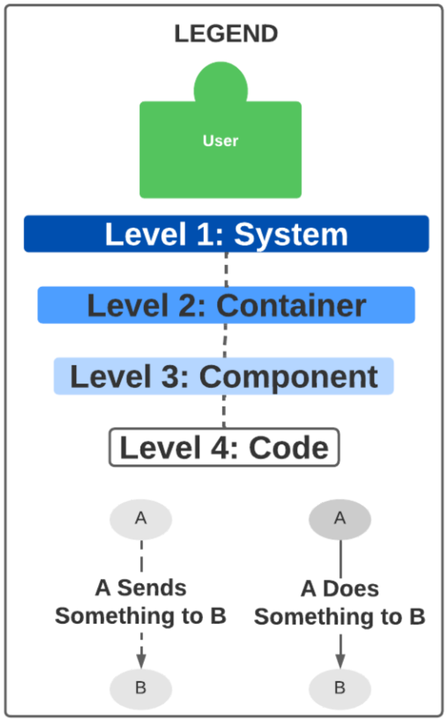
Mission Measurement Data (not simulated data, synthetic data, etc) does not come to the DAAC until Phase E (after launch).
Many MDPS are not sensor/platform-based processing that shares similar architecture. e.g. HLS, SNWG/OPERA, and MEaSUREs class projects.

The MDPS is tuned for Project Users. General Scientists do not have access to the MDPS.

The MDPS may use on-prem, cloud, or hybrid infrastructure.

The MDPS does not have any dependency on any external systems such as the DAAC.

Some issues with these systems: duplication of effort in each instance (development, maintanance, operations, etc), not open science friendly as there is no General Scientist access, is not tuned for System Science (access to data and algorithms from multiple missions).
SPD-41 section IV moving forward prohibits data embargos.



LEGEND

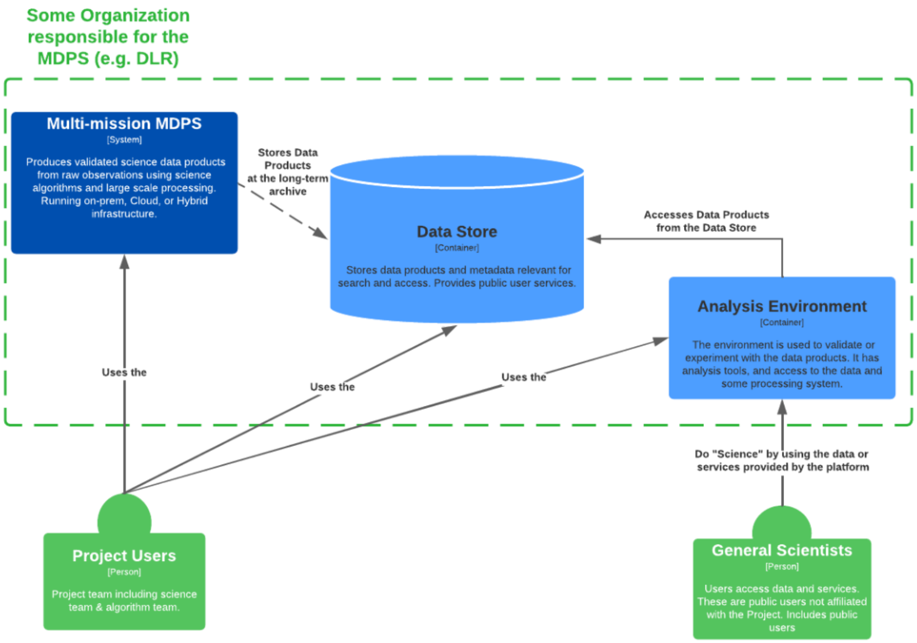
User

Level 1: System

Level 2: Container

Level 3: Component

Level 4: Code


Some Organization responsible for the MDPS (e.g. DLR)
Multi-mission MDPS
[System]
Stores Data Products at the long-term archive
Data Store
[Container]
Accesses Data Products from the Data Store
Analysis Environment
[Container]
Uses the
Uses the
Uses the
Do "Science" by using the data or services provided by the platform
Project Users
[Person]
General Scientists
[Person]

Produces validated science data products from raw observations using science algorithms and large scale processing. Running on-prem, Cloud, or Hybrid infrastructure.

Stores data products and metadata relevant for search and access. Provides public user services.

The environment is used to validate or experiment with the data products. It has analysis tools, and access to the data and some processing system.

Project team including science team & algorithm team.

Users access data and services. These are public users not affiliated with the Project. Includes public users

Architecture 2 - The Multi-mission System (e.g. DLR, NOA, IEMGO (isro))

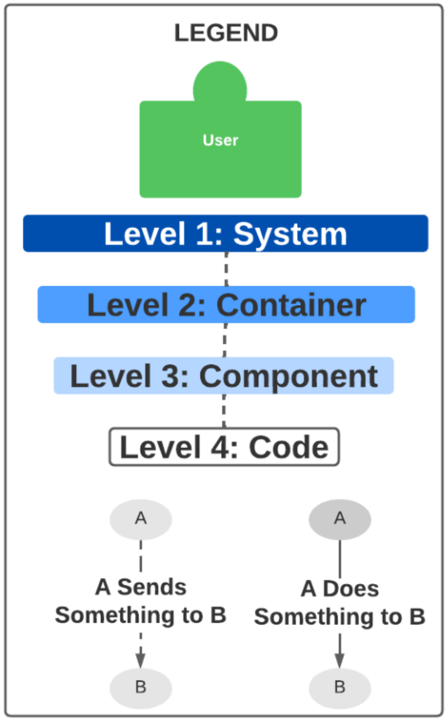
One instance of a multi-mission MDPS that processes data from multiple missions

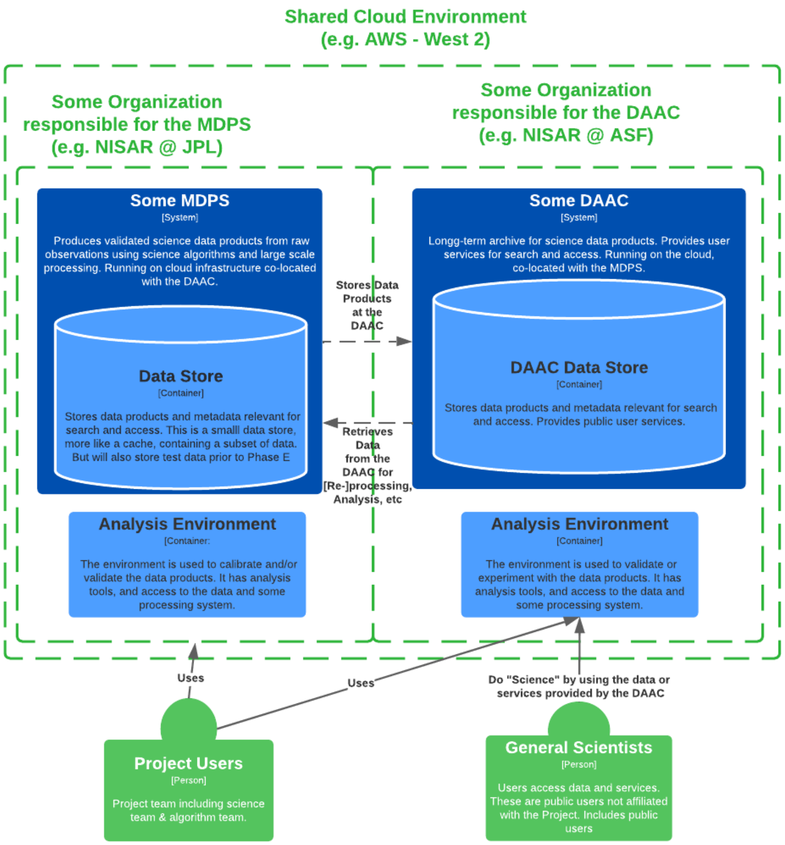
The long term archive may be internal to the organization or outside, but it's easily accessible (like in the case of DLR's terrabyte system) even if there is an external organization responsible for it.

The additional component that seems to stand out in this architecture that was not a highlight in Architecture 1 is the Analysis Environment. The multi-mission nature of this architecture (having all the data and algorithms) lends itself well to that.

The MDPS may utilize on-prem, cloud, or hybrid infrastructure

Some issues with these systems: cost sharing, coupling failures, dependency across missions - No software Freezes. This architecture seems to lend itself well when it's funded/maintained within one organization.





Shared Cloud Environment
(e.g. AWS - West 2)

Some Organization responsible for the MDPS
(e.g. NISAR @ JPL)

Some Organization responsible for the DAAC
(e.g. NISAR @ ASF)

Some MDPS
[System]
Produces validated science data products from raw observations using science algorithms and large scale processing. Running on cloud infrastructure co-located with the DAAC.

Data Store
[Container]
Stores data products and metadata relevant for search and access. This is a smalll data store, more like a cache, containing a subset of data. But will also store test data prior to Phase E

Some DAAC
[System]
Longg-term archive for science data products. Provides user services for search and access. Running on the cloud, co-located with the MDPS.

DAAC Data Store
[Container]
Stores data products and metadata relevant for search and access. Provides public user services.

Stores Data Products at the DAAC

Retrieves Data from the DAAC for [Re-]processing, Analysis, etc

Analysis Environment
[Container:
The environment is used to calibrate and/or validate the data products. It has analysis tools, and access to the data and some processing system.

Analysis Environment
[Container]
The environment is used to validate or experiment with the data products. It has analysis tools, and access to the data and some processing system.

Uses

Uses

Do "Science" by using the data or services provided by the DAAC

Project Users
[Person]
Project team including science team & algorithm team.

General Scientists
[Person]
Users access data and services. These are public users not affiliated with the Project. Includes public users

Architecture 3 - Co-located MDPS & DAAC (e.g.

Each MDPS is a separate instance independent of other MDPS's the organization may implement

Each instance of an MDPS mostly uses common components and common open source technologies, but those are generally only shared within one organization (i.e. Goddard, JPL, NOA, USGS, DLR, etc. but not shared across organizations)

The MDPS maintains only a small subset of the data, utilizing the DAAC storage both as the life of mission and long-term archive

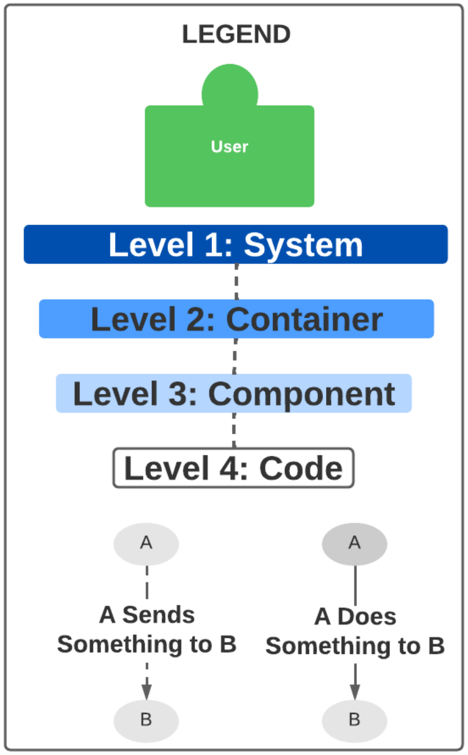
Data is available at the DAAC as soon as it's produced at the MDPS, even at early phases of the mission (pre-launch test data for example).

The MDPS and DAAC are co-located in the Cloud

The MDPS is tuned for Project Users. General Scientists do not have access to the MDPS.

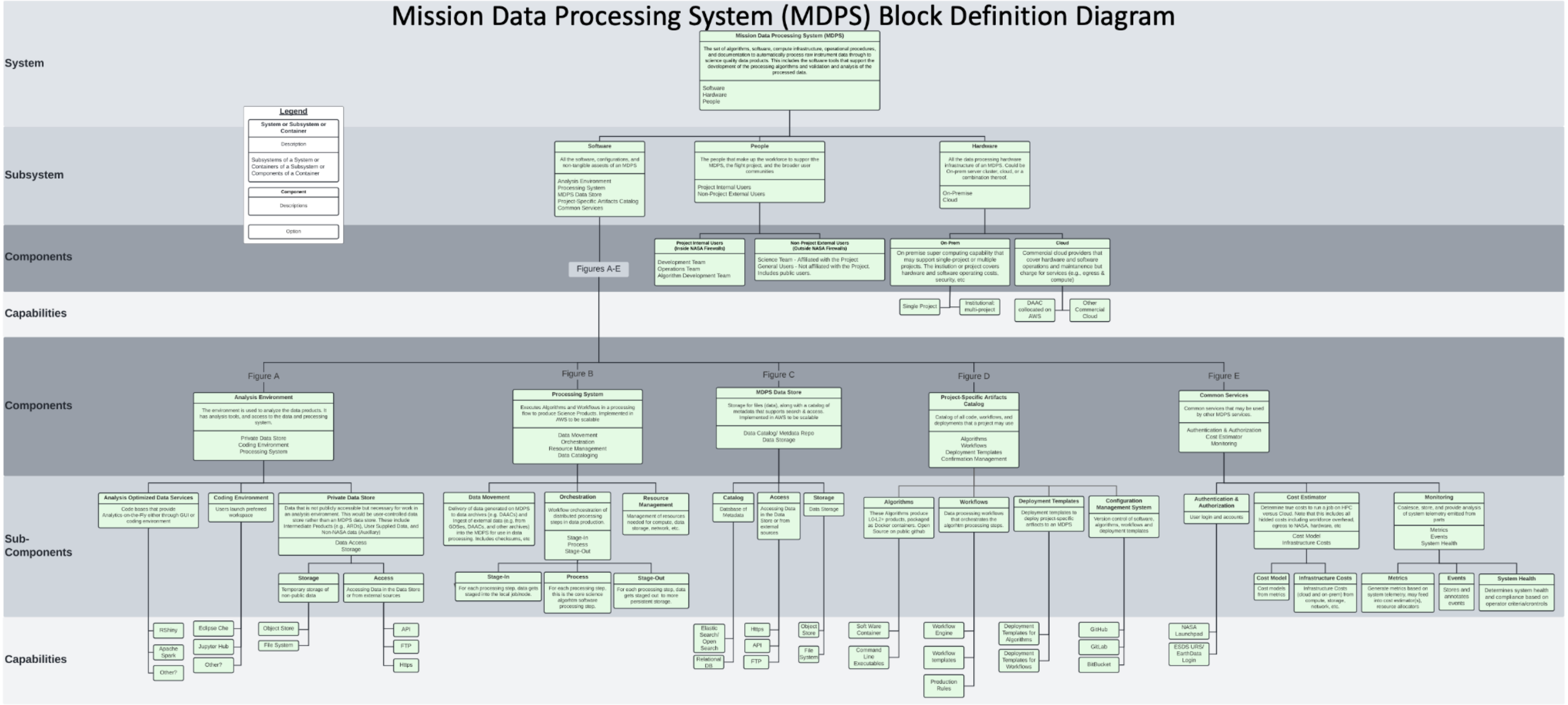
The MDPS does not have any dependency on any external systems such as the DAAC.

The MDPS can access data from DAACs for bulk reprocessing.

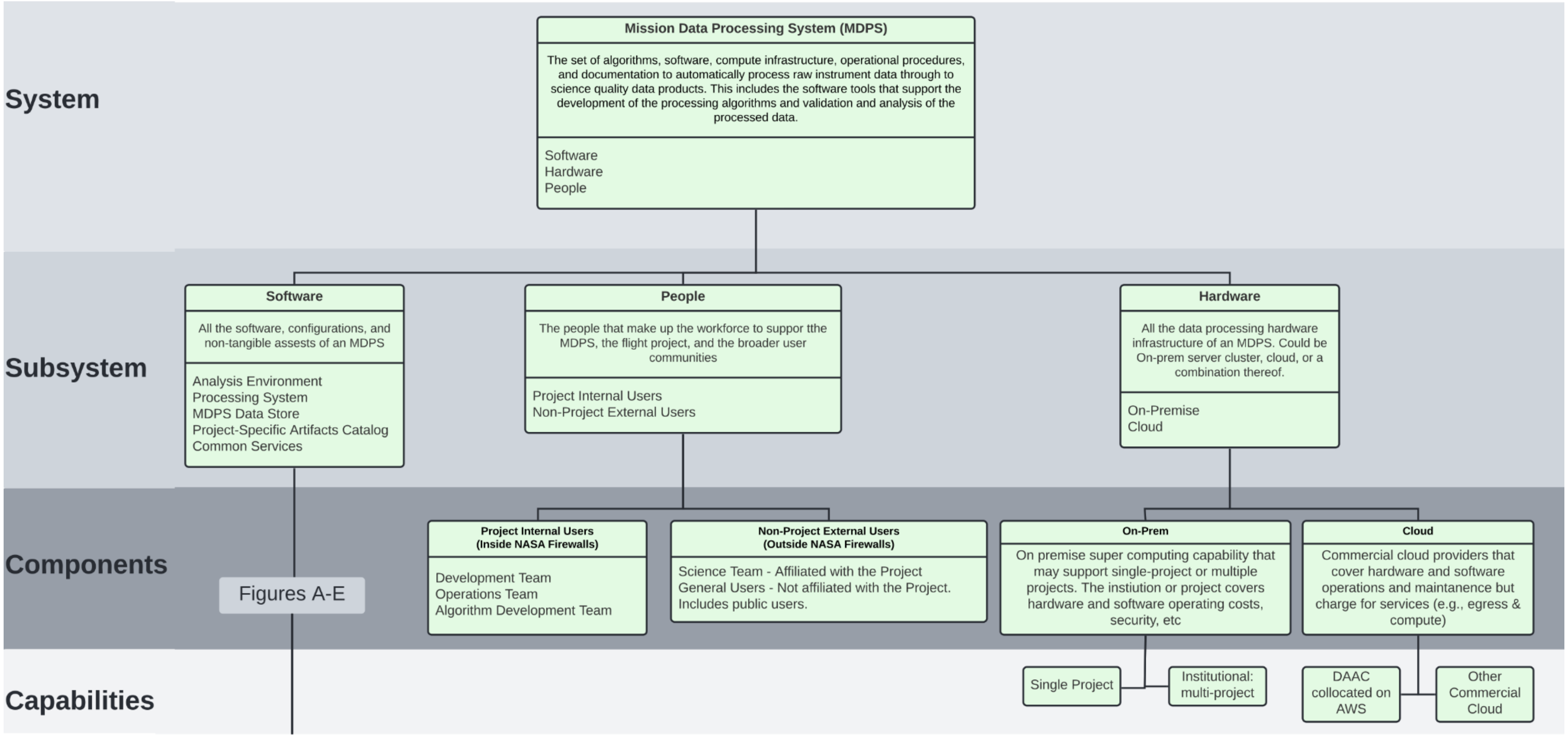


7.4.2 MDPS Block Definition Diagram

A Block Definition Diagram (BDD) is used to define a system, its subsystems, components, and sub-components. Based on a synthesis of Mission Data Processing Systems (MDPS) presented at this workshop, the SAWG developed the following BDD:


Mission Data Processing System (MDPS) Block Definition Diagram
System
Subsystem
Components
Capabilities
Components
Sub-Components
Capabilities
Legend
System or Subsystem or Container
Description
Subsystems of a System or Containers of a Subsystem or Components of a Container
Component
Descriptions
Option
Mission Data Processing System (MDPS)
The set of algorithms, software, compute infrastructure, operational procedures, and documentation to automatically process raw instrument data through to science quality data products. This includes the software tools that support the development of the processing algorithms and validation and analysis of the processed data.
Software
Hardware
People
Software
All the software, configurations, and non-tangible assests of an MDPS
Analysis Environment
Processing System
MDPS Data Store
Project-Specific Artifacts Catalog
Common Services
People
The people that make up the workforce to suppor tthe MDPS, the flight project, and the broader user communities
Project Internal Users
Non-Project External Users
Hardware
All the data processing hardware infrastructure of an MDPS. Could be On-prem server cluster, cloud, or a combination thereof.
On-Premise
Cloud
Project Internal Users (Inside NASA Firewalls)
Development Team
Operations Team
Algorithm Development Team
Non-Project External Users (Outside NASA Firewalls)
Science Team - Affiliated with the Project
General Users - Not affiliated with the Project. Includes public users.
On-Prem
On premise super computing capability that may support single-project or multiple projects. The instiution or project covers hardware and software operating costs, security, etc
Cloud
Commercial cloud providers that cover hardware and software operations and maintanence but charge for services (e.g., egress & compute)
Single Project
Institutional: multi-project
DAAC collocated on AWS
Other Commercial Cloud
Figures A-E
Figure A
Analysis Environment
The environment is used to analyze the data products. It has analysis tools, and access to the data and processing system.
Private Data Store
Coding Environment
Processing System
Figure B
Processing System
Executes Algorithms and Workflows in a processing flow to produce Science Products. Implemented in AWS to be scalable
Data Movement
Orchestration
Resource Management
Data Cataloging
Figure C
MDPS Data Store
Storage for files (data), along with a catalog of metadata that supports search & access. Implemented in AWS to be scalable
Data Catalog/ Metdata Repo
Data Storage
Figure D
Project-Specific Artifacts Catalog
Catalog of all code, workflows, and deployments that a project may use
Algorithms
Workflows
Deployment Templates
Confirmation Management
Figure E
Common Services
Common services that may be used by other MDPS services.
Authentication & Authorization
Cost Estimator
Monitoring
Analysis Optimized Data Services
Code bases that provide Analytics-on-the-Fly either through GUI or coding environment
Coding Environment
Users launch preferred workspace
Private Data Store
Data that is not publicly accessible but necessary for work in an analysis environment. This would be user-controlled data store rather than an MDPS data store. These include Intermediate Products (e.g., ARDs), User Supplied Data, and Non-NASA data (Auxiliary)
Data Access
Storage
Storage
Temporary storage of non-public data
Access
Accessing Data in the Data Store or from external sources
RShiny
Apache Spark
Other?
Eclipse Che
Jupyter Hub
Other?
Object Store
File System
API
FTP
Https
Data Movement
Delivery of data generated on MDPS to data archives (e.g. DAACs) and Ingest of external data (e.g. from GDSes, DAACs, and other archives) into the MDPS for use in data processing. Includes checksums, etc
Orchestration
Workflow orchestration of distributed processing steps in data production.
Stage-In
Process
Stage-Out
Resource Management
Management of resources needed for compute, data storage, network, etc.
Stage-In
For each processing step, data gets staged into the local job/node.
Process
For each processing step, this is the core science algorihm software processing step.
Stage-Out
For each processing step, data gets staged out to more persistent storage.
Catalog
Database of Metadata
Access
Accessing Data in the Data Store or from external sources
Storage
Data Storage
Elastic Search/ Open Search
Relational DB
Https
API
FTP
Object Store
File System
Algorithms
These Algorithms produce L0-L2+ products, packaged as Docker containers. Open Source on public github
Workflows
Data processing workflows that orchestrates the algorihm processing steps.
Deployment Templates
Deployment templates to deploy project-specific artifacts to an MDPS
Configuration Management System
Version control of software, algorithms, workflows and deployment templates
Soft Ware Container
Command Line Executables
Workflow Engine
Workflow templates
Production Rules
Deployment Templates for Algorithms
Deployment Templates for Workflows
GitHub
GitLab
BitBucket
Authentication & Authorization
User login and accounts
NASA Launchpad
ESDS URS/ EarthData Login
Cost Estimator
Determine true costs to run a job on HPC versus Cloud. Note that this includes all hidded costs including workforce overhead, egress to NASA, hardware, etc
Cost Model
Infrastructure Costs
Cost Model
Cost models from metrics
Infrastructure Costs
Infrastructure Costs (cloud and on-prem) from compute, storage, network, etc.
Monitoring
Coalesce, store, and provide analysis of system telemetry emitted from parts
Metrics
Events
System Health
Metrics
Generate metrics based on system telemetry, may feed into cost estimator(s), resource allocators
Events
Stores and annotates events
System Health
Determines system health and compliance based on operator criteria/controls

In the following figures we will zoom in to show the descriptions of first the system and its subsystems:



Mission Data Processing System (MDPS)

The set of algorithms, software, compute infrastructure, operational procedures, and documentation to automatically process raw instrument data through to science quality data products. This includes the software tools that support the development of the processing algorithms and validation and analysis of the processed data.

Software
Hardware
People

Software

All the software, configurations, and non-tangible assests of an MDPS

Analysis Environment
Processing System
MDPS Data Store
Project-Specific Artifacts Catalog
Common Services

People

The people that make up the workforce to suppor tthe MDPS, the flight project, and the broader user communities

Project Internal Users
Non-Project External Users

Hardware

All the data processing hardware infrastructure of an MDPS. Could be On-prem server cluster, cloud, or a combination thereof.

On-Premise
Cloud

Project Internal Users (Inside NASA Firewalls)

Development Team
Operations Team
Algorithm Development Team

Non-Project External Users (Outside NASA Firewalls)

Science Team - Affiliated with the Project
General Users - Not affiliated with the Project. Includes public users.

On-Prem

On premise super computing capability that may support single-project or multiple projects. The instiution or project covers hardware and software operating costs, security, etc

Cloud

Commercial cloud providers that cover hardware and software operations and maintanence but charge for services (e.g., egress & compute)

Then of the software subsystem component A:

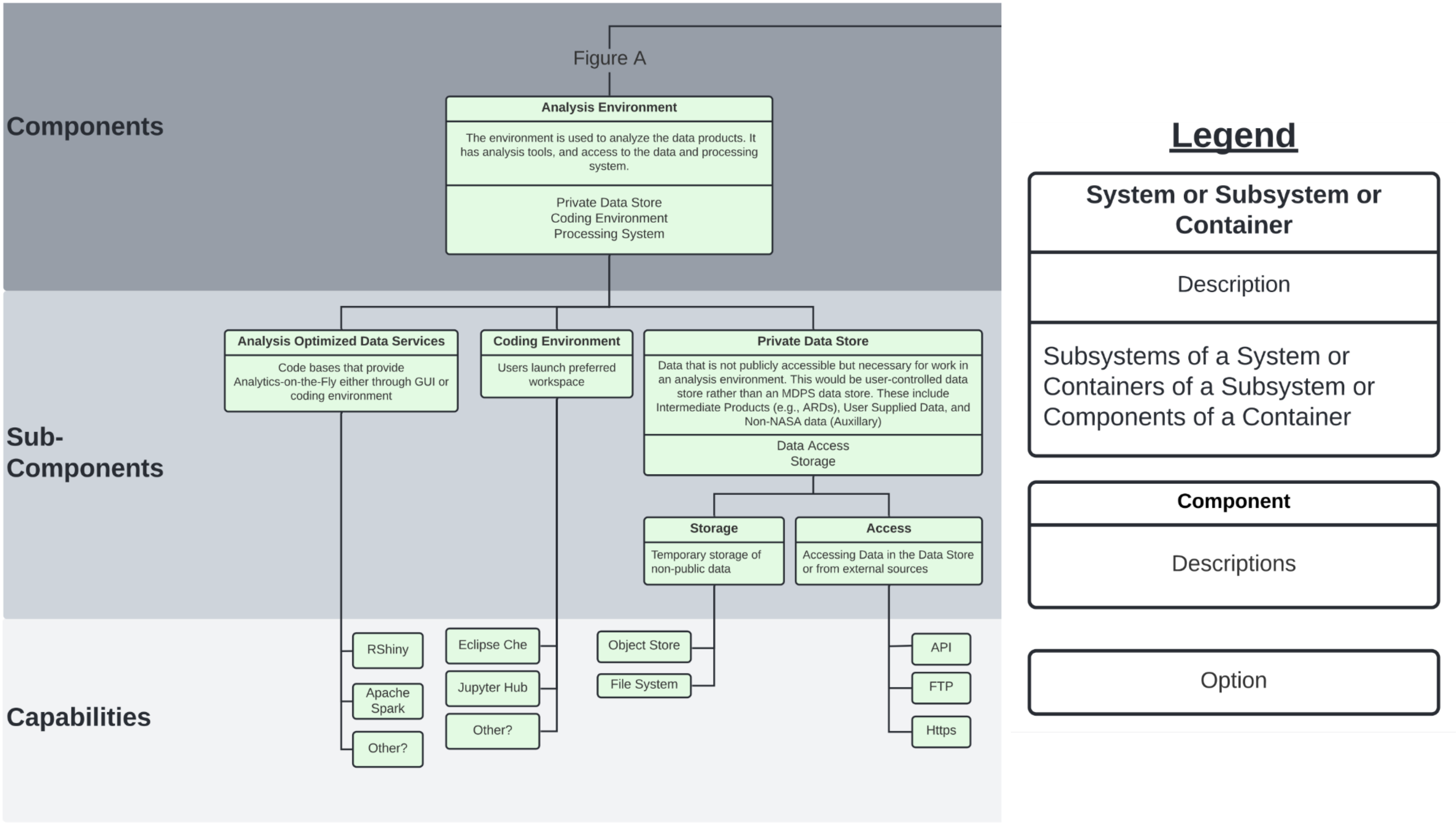


Figure A

Components

Analysis Environment

The environment is used to analyze the data products. It has analysis tools, and access to the data and processing system.

Private Data Store
Coding Environment
Processing System

Sub-Components

Analysis Optimized Data Services

Code bases that provide Analytics-on-the-Fly either through GUI or coding environment

Coding Environment

Users launch preferred workspace

Private Data Store

Data that is not publicly accessible but necessary for work in an analysis environment. This would be user-controlled data store rather than an MDPS data store. These include Intermediate Products (e.g., ARDs), User Supplied Data, and Non-NASA data (Auxillary)

Data Access
Storage

Storage

Temporary storage of non-public data

Access

Accessing Data in the Data Store or from external sources

Capabilities

RShiny
Apache Spark
Other?

Eclipse Che
Jupyter Hub
Other?

Object Store
File System

API
FTP
Https

Legend

System or Subsystem or Container

Description

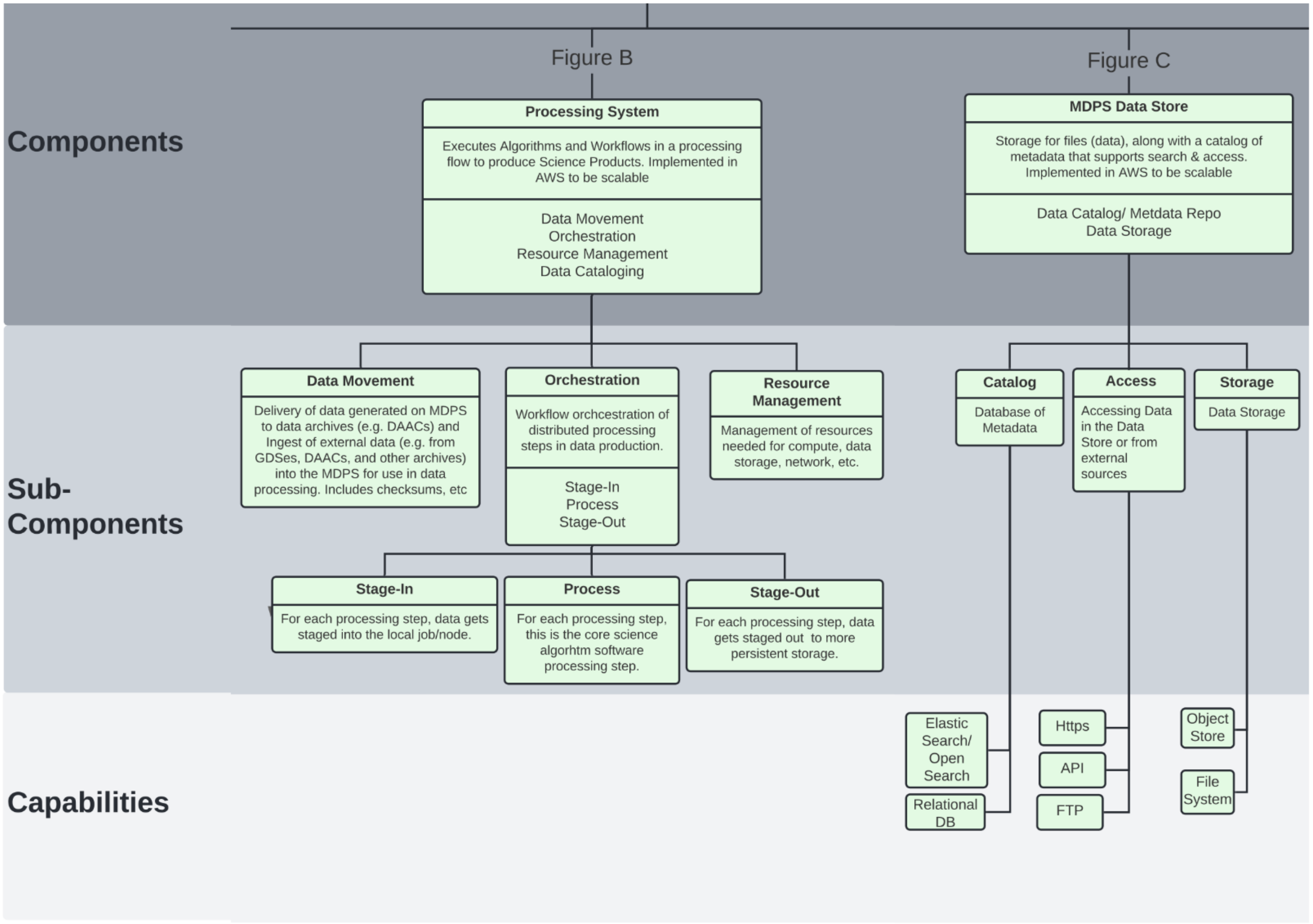
Subsystems of a System or Containers of a Subsystem or Components of a Container

Component

Descriptions

Option

Then software subsystem components B and C:



Processing System

Executes Algorithms and Workflows in a processing flow to produce Science Products. Implemented in AWS to be scalable

Data Movement
Orchestration
Resource Management
Data Cataloging

MDPS Data Store

Storage for files (data), along with a catalog of metadata that supports search & access. Implemented in AWS to be scalable

Data Catalog/ Metdata Repo
Data Storage

Data Movement

Delivery of data generated on MDPS to data archives (e.g. DAACs) and Ingest of external data (e.g. from GDSes, DAACs, and other archives) into the MDPS for use in data processing. Includes checksums, etc

Orchestration

Workflow orchcestration of distributed processing steps in data production.

Stage-In
Process
Stage-Out

Resource Management

Management of resources needed for compute, data storage, network, etc.

Catalog

Database of Metadata

Access

Accessing Data in the Data Store or from external sources

Storage

Data Storage

Stage-In

For each processing step, data gets staged into the local job/node.

Process

For each processing step, this is the core science algorhtm software processing step.

Stage-Out

For each processing step, data gets staged out to more persistent storage.

Then software subsystem component D:

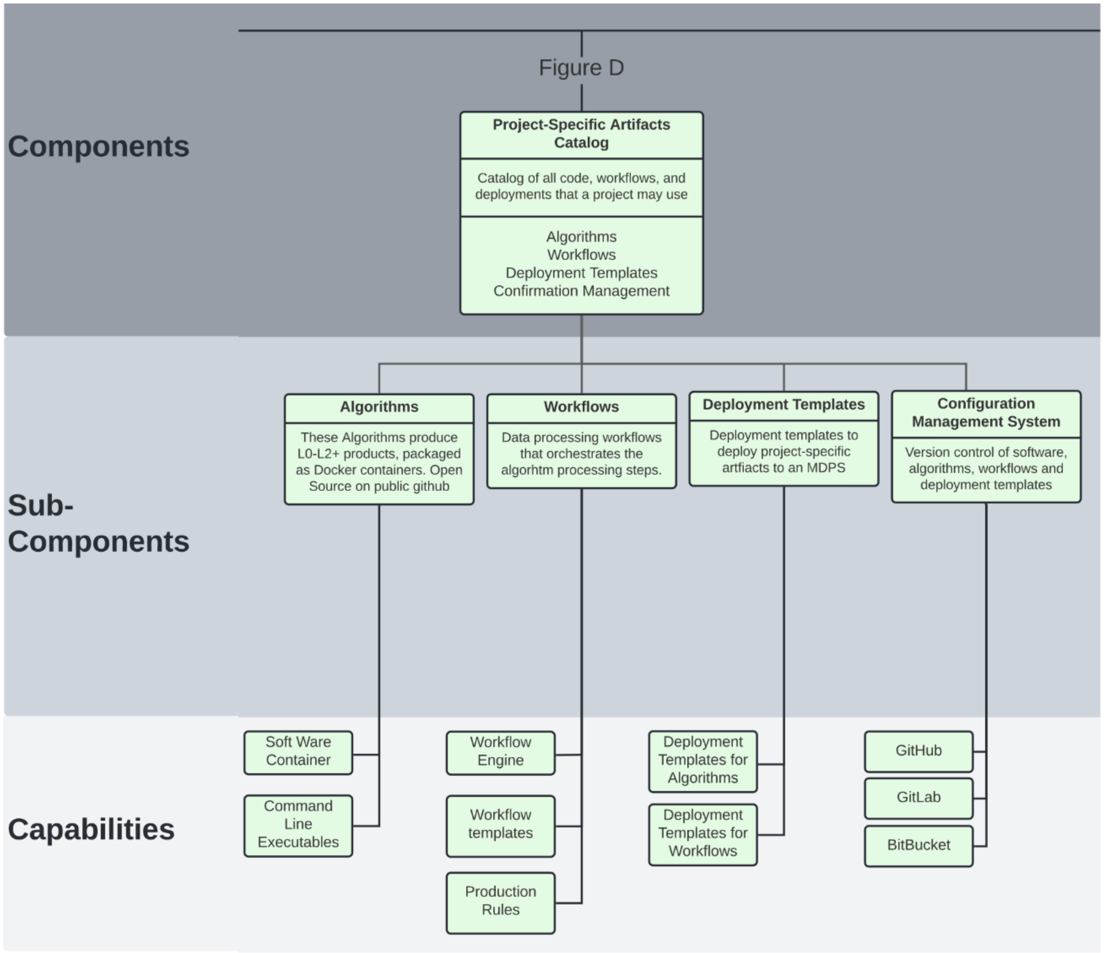


Figure D
Components
Project-Specific Artifacts Catalog
Catalog of all code, workflows, and deployments that a project may use
Algorithms
Workflows
Deployment Templates
Confirmation Management
Sub-Components
Algorithms
These Algorithms produce L0-L2+ products, packaged as Docker containers. Open Source on public github
Workflows
Data processing workflows that orchestrates the algorhtm processing steps.
Deployment Templates
Deployment templates to deploy project-specific artfiacts to an MDPS
Configuration Management System
Version control of software, algorithms, workflows and deployment templates
Capabilities
Soft Ware Container
Command Line Executables
Workflow Engine
Workflow templates
Production Rules
Deployment Templates for Algorithms
Deployment Templates for Workflows
GitHub
GitLab
BitBucket

Finally, the software subsystem component E:

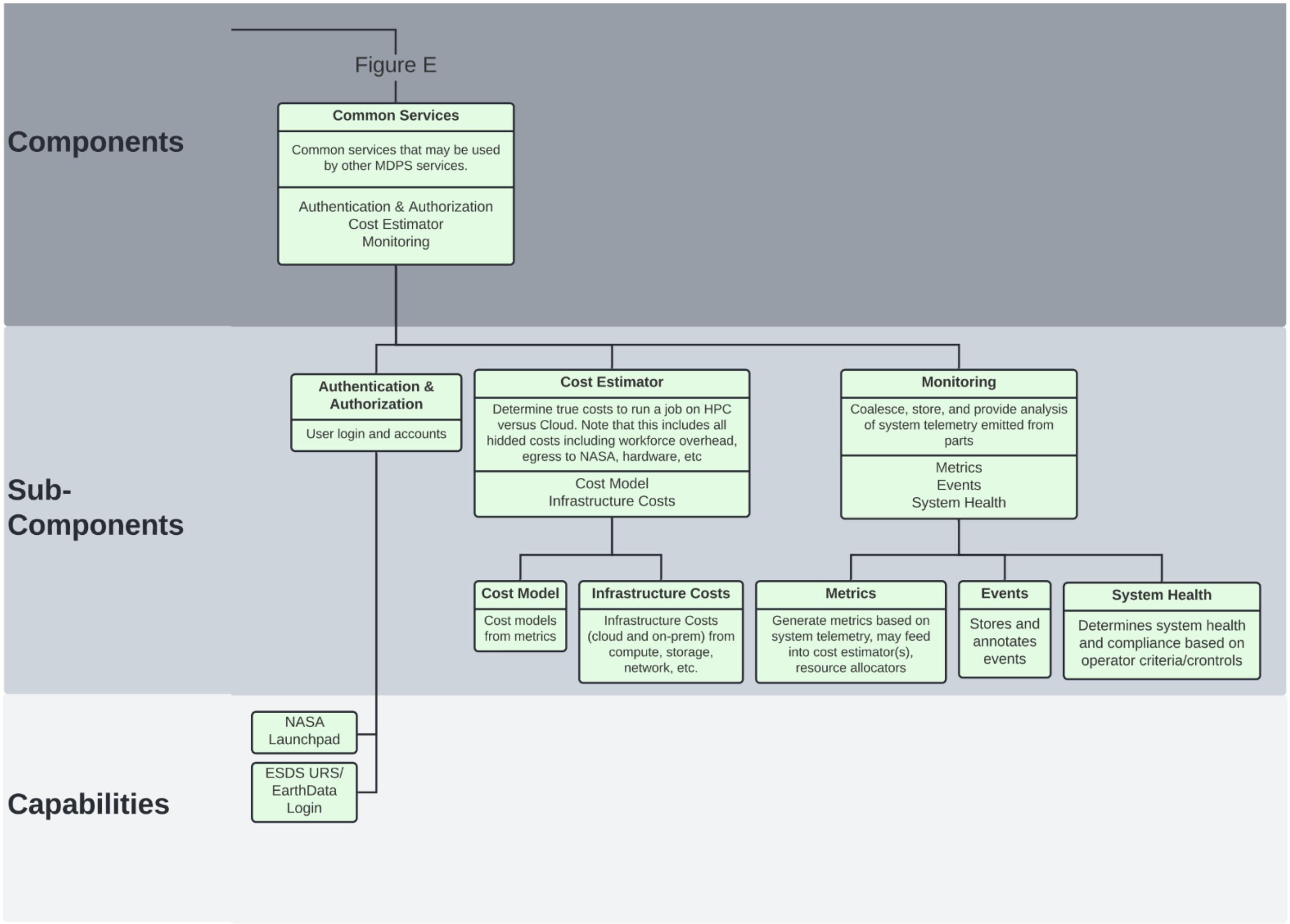


Figure E
Components
Sub-Components
Capabilities
Common Services
Common services that may be used by other MDPS services.
Authentication & Authorization
Cost Estimator
Monitoring
Authentication & Authorization
User login and accounts
Cost Estimator
Determine true costs to run a job on HPC versus Cloud. Note that this includes all hidded costs including workforce overhead, egress to NASA, hardware, etc
Cost Model
Infrastructure Costs
Monitoring
Coalesce, store, and provide analysis of system telemetry emitted from parts
Metrics
Events
System Health
Cost Model
Cost models from metrics
Infrastructure Costs
Infrastructure Costs (cloud and on-prem) from compute, storage, network, etc.
Metrics
Generate metrics based on system telemetry, may feed into cost estimator(s), resource allocators
Events
Stores and annotates events
System Health
Determines system health and compliance based on operator criteria/crontrols
NASA Launchpad
ESDS URS/ EarthData Login

7.5 Path Forward

The results from this workshop in conjunction with workshop 1 on evaluation criteria for a data system that meets the four study objectives (Section 5) will be used to conduct a trade study. We will expand upon the three common architectures (Section 7.4.1) to include additional architectures using both C4 models and deployment views that include the subsystems, components, and subcomponents of an MDPS (Section 7.4.2). We will then use a summary matrix (Felix, 2004; NASA, 2007 Section 6.8 Decision Analysis) that evaluates each architecture against our evaluation criteria. Each criteria needs to be independent and discriminating. They can be hierarchical (tier 1 vs tier 2) and provide a priority/weighting. The SAWG will meet with the Steering Committee to determine a process for assigning weightings. Cost and risk analysis will be post-analysis. The final recommendation will include both quantitative assessment and qualitative considerations of costs and risks associated with maturity of different approaches and feasibility for meeting ESO schedules and cost constraints.

References

Carroll, S. R., Garba, I., Figueroa-Rodriguez, O. L., Halbrook, J., Raseroka, K., Rodriguez-Lonebear, D., et al. (2020). The CARE Principles for Indigenous Data Governance. Retrieved April 20, 2021, from https://datascience.codata.org/articles/10.5334/dsj-2020-043/

Felix, A. (2004). Standard Approach to Trade Studies: A Process Improvement Model that Enables Systems Engineers to Provide Information to the Project Manager by Going Beyond the Summary Matrix. Presented at the International Council on Systems Engineering (INCOSE) Mid-Atlantic Regional Conference. Retrieved from https://www.acqnotes.com/Attachments/Standard%20Approach%20to%20Trade%20Studies.pdf

Hood, R., & Jin, H. (n.d.). What role should commercial clouds play in NASA HPC? Retrieved April 22, 2022, from https://www.nas.nasa.gov/upload/files/sc18posters/5_Hood_R_SC18_CloudStudy_LR.pdf

Ivezić, Ž., Kahn, S. M., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., et al. (2019). LSST: From Science Drivers to Reference Design and Anticipated Data Products. *The Astrophysical Journal*, *873*(2), 111. https://doi.org/10.3847/1538-4357/ab042c

Margetta, R. (2021, May 24). New NASA Earth System Observatory to Help Address Climate Change [Text]. Retrieved June 7, 2021, from http://www.nasa.gov/press-release/new-nasa-earth-system-observatory-to-help-address-mitigate-climate-change

NASA. (2007). NASA Systems Engineering Handbook. *NASA SP-2016-6105*, 297. Retrieved from https://www.nasa.gov/seh/1-introduction

NASA. (2019). *Strategy for Data Management and Computing for Groundbreaking Science 2019-2024*. Retrieved from https://go.nasa.gov/3HzilNg

NASA. (2021). *SMD Policy Document SPD-41*. Retrieved from https://science.nasa.gov/science-red/s3fs-public/atoms/files/Scientific%20Information%20policy%20SPD-41.pdf

NRC. (2019). *Thriving on Our Changing Planet: A Decadal Strategy for Earth Observation from Space: An Overview for Decision Makers and the Public* (p. 25437). Washington, D.C.: National Academies Press. https://doi.org/10.17226/25437

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Glossary

Accessible - Data, tools, software, documentation, publications follow FAIR Data Principles.

Analysis-Ready Data - 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.

Application - use of NASA data for decision support (policy, resources, etc).

Analysis-Ready Cloud Optimized (ARCO) - ARD data stored in cloud-optimized data formats enabling rapid access to the ARDs.

Application-Ready Data: GIS-ready data

Architecture: A MDPS architecture is a system as a collection of components and connectors. Architecture should not be considered merely a set of models or structures, but should include the decisions that lead to these particular structures, and the rationale behind them.

Baseline Architecture - The best architecture that meets Workshop 1 evaluation criteria and is additive to a Threshold Architecture, such that should budget run over, components could be descoped without compromising our ability to meet the evaluation criteria.

Benchmark Architecture - a reference architecture for the current implementation.

Cloud - Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. (NIST SP 800-145, 2011)

Capability Need - functionalities of the system.

Centralized Service - A common instance of a running system shared by many users.

Common Capability - Common, shared software that is deployed separately by different users.

Common ESO data system - a standardized MDPS that services multiple ESO missions.

Common Service - a service that is needed by many projects but may be implemented, deployed independently by each project.

Data Active Archive Centers (DAACs) - are NASA data archives that serve different research communities but share common services to standardize NASA data management and archive through the NASA Earth Science Data and Information System (ESDIS).

Data Lake - The concept of data-proximate processing where the data is stored and co-located with the processing.

Data Product Level - All definitions are assumed to be consistent with the NASA Data Processing Levels: https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy/data-levels

Earth System Observatory (ESO) - a constellation of satellites will be launched by NASA in the 2020s to observe the Earth System as designated by the National Academies Decadal Survey (NRC, 2019) and classified as “designated observables”.

Evaluation Criteria - are design constraints by which to evaluate different architectures. This term is used in place of “requirements”, which are often traced for data systems from higher-level mission requirements; hence the avoidance of prescribing them for all ESO and future missions.

FAIR Data Principles - Data should be Findable, Accessible, Interoperable, and Reproducible by machines (Wilkinson et al., 2016).

Federated services - a service that is owned and operated by one organization, but is contributed to by many projects.

Ground Data System - The system responsible for receiving telemetry data from the observatory and providing it to the MDPS, which does the instrument specific processing.

Hybrid Cloud - Infrastructure that is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability. (NIST SP 800-145, 2011)

Inclusive - The process and participants welcome participation by and collaboration with diverse people and organizations.

Latency - defined as time between acquisition and data access by the users.

Mission Data Processing System (MDPS) - The set of algorithms, software, compute infrastructure, operational procedures, and documentation to automatically process raw instrument data through to science quality data products. This includes the software tools that support the development of the processing algorithms and validation and analysis of the processed data.

On-prem Computing - Computing infrastructure that physically resides within an enterprise owned data center, server room, etc. On-prem may be referred to as “in-house”. Usually, an organization is fully responsible for procuring, deploying and managing on-prem computing.

Open Science - “a collaborative culture enabled by technology that empowers the open sharing of data, information, and knowledge within the scientific community and the wider public to accelerate scientific research and understanding” (Ramachandran et al., 2021).

Open Source Science - builds on concepts from Open Source Software revolution that expanded participation in developing code and applies it to the scientific process to accelerate discovery through open science from project initiation through implementation.

Open Source Software - The Open Source Initiative (OSI) defines Software to be Open Source if distributed under a license with a set of criteria: 1) license shall not restrict any party from selling or giving away the software, i.e. free redistribution, 2) source code is included with any program or set of programs, 3) license allows for derived works, 4) integrity of author's source code, 5) licence must not discriminate against a groups or persons, 6) license must not discrimination against fields of endeavor, 7) any rights must apply to all whom a program or source is redistributed to, 8) rights attached to the program must not depend on the program's being part of a particular software distribution, 9) license must not place restrictions on other software that is distributed along with, and 10) no provision of the license may be predicated on any individual technology or style of interface. (https://opensource.org/osd)

Permissive software - software that can be copied, modified, redistributed, etc.

Reproducible - The scientific process and results can be reproduced by members of the community.

Scientific Information - publications, data, and software

Shared services - a service owned and managed by one organization that is used by many projects.

System Architecture Working Group (SAWG) - a team of system engineers, data system architects, software engineers, and ESO mission representatives tasked with conducting the ESO open source science data system architecture study. The SAWG is composed of science data system experts who represent the diversity of the data system community and are connected to the end-user science community and the ESO missions.

Steering Committee - the leadership team for the ESO open source science data system architecture study responsible for providing programmatic insights and steering the SAWG to conduct a programmatically relevant study.

Threshold Architecture - The bare minimum architecture needed to meet the evaluation criteria from Workshop 1.

Transparency - Both the scientific process and results are visible, accessible and understandable.

Acronyms

ACCP	Aerosol, Cloud, Convection, and Precipitation
ACCESS	Advancing Collaborative Connections for Earth System Science
ACF	Analytic Center Frameworks
ADE	Application Development Environment
AGU	American Geophysical Union
AI	Artificial Intelligence
AIRS	Atmospheric Infrared Sounder
AIST	Advanced Information Systems Technology
ALIAS	Automated Labeling for Interactive Assisted Segmentation
AMS	American Meteorological Society
AMMOS	Advanced Multi-Mission Operations System
ACF	Analytics Collaborative Framework
AOS	Atmosphere Observing System
API	Application Programming Interface
ARCO	Analysis-Ready Cloud-Optimized data
ARD	Analysis-Ready Data
ARSET	Applied Remote Sensing Training
ASF	Alaska Satellite Facility
ASI	Agenzia Spaziale Italiana
ASP	Applied Sciences Program
ATBD	Algorithm Theoretical Basis Documents
ATLAS	Advanced Topographic Laser Altimeter System
AWS	Amazon Web Services
BDD	Block Definition Diagram

cal/val	Calibration and validation
CCAP	Containerized Cloud Algorithm Package
CCSDS	Consultative Community for Science Data Systems
CERES	Cloud and the Earth's Radiant Energy System
CHIME	Canadian Hydrogen Intensity Mapping Experiment
CI/CD	Continuous Integration/Continuous Delivery
CLARREO	CLimate Absolute Radiance REfractivity Observatory
CMR	Common Metadata Repository
CNES	Centre National d'Etudes Spatiales
COG	Cloud-Optimized GeoTIFF
CSA	Canadian Space Agency
CSP	Cloud Service Provider
DAAC	Distributed Active Archive Center
DB	Data Base
DEVELOP	Digital Earth Virtual Environment and Learning Outreach Program
DMP	Data Management Plan
DMS	Data Management System
DOI	Digital Object Identifier
DPC	Data Processing Center
DPS	Data Processing Service
EDC	Earthdata cloud
EDOS	Earth Orbiting System (EOS) Data and Operations System
EIS	Earth Information System
EMIT	Earth Surface Mineral Dust Source Investigation

EMS	EOSDIS Metrics System
EO	Earth Observation
EOS	Earth Orbiting System
EOSDIS	Earth Observing System Data and Information System
ESA	European Space Agency
ESD	Earth Science Division
ESDIS	Earth Science Data and Information System
ESIP	Earth Science Information Partners
ESTO	Earth Science Technology Office
ESO	Earth System Observatory
ETL	Extraction, Transformation, and Loading
FAIR	Findable, Accessible, Inter-operable, Reproducible
FORGE	Future Operationally Resilient Ground Evolution
GDS	Ground Data System
GEDI	Global Ecosystem Dynamics Investigation
GEE	Google Earth Engine
GES-DISC	Goddard Earth Sciences Data and Information Service Center
GFO	Gravity Recovery and Climate Experiment Follow-On
GFZ	Geoforschungszentrum
GIBS	Global Imagery Browse Services
GIS	Geographic Information System
GNSS	Global Navigation Satellite System
GNU	GNU's Not Unix
GPU	Graphics Processing Unit

GRACE	Gravity Recovery and Climate Experiment
GRACE-FO	Gravity Recovery and Climate Experiment Follow-On
GSFC	Goddard Space Flight Center
GUI	Graphical User Interface
HARP	Hyper Angular Rainbow Polarimeter
HCSA	Hybrid Computing and Storage Architectures
HEC	High-end Computing
HECC	High-end Computing Capability
HOSC	Huntsville Operations Support Center
HPC	High Performance Computing
HQ	Headquarters
HyP3	Hybrid Pluggable Processing Pipeline
IDE	Integrated Development Environment
IMGEOS	Integrated Multi-Mission Ground Segment for Earth Observing Satellites
InSAR	Interferometric Synthetic Aperture Radar
IP	Internet Protocol
IRAD	Internal Research and Development
ISS	International Space Station
ISRO	Indian Space Research Organisation
ISCE	InSAR Scientific Computing Environment
ITAR	International Traffic in Arms Regulations
JAXA	Japan Aerospace Exploration Agency
JDK	Java Development Kit (Open JDK)
JEM-EF	Japanese External Module- Exposed Facility

JPL	Jet Propulsion Laboratory
JPSS	Joint Polar Satellite System
JSON	JavaScript Object Notation
KVM	Kernel based Virtual Machine
L#	Data Product Level # as defined by https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy/data-levels
LAADS	Land And Atmosphere Distribution System
LANCE	Land Atmosphere Near real-time Capability for EOS
LC	LandSat Cloud
LEO	Low Earth Orbit
LIDAR	Light Detection and Ranging
LRZ	Leibniz Supercomputing Centre
LSST	Legacy Survey of Space and Time
LSTM	Long Short Term Memory
MAAP	Multi-Mission Algorithm and Analysis Platform
MADS	Mission Access Data System
MAIA	Multi-Angle Imager for Aerosols
MC	Mass Change
MCP	Microsoft Cloud Platform
MDPAF	Mission Data Processing Application Framework
MDPS	Mission Data Processing System
MOC	Mission Operations Center
MODAPS	MODIS Adaptive Processing System
MODIS	MODerate resolution Imaging Spectrometer

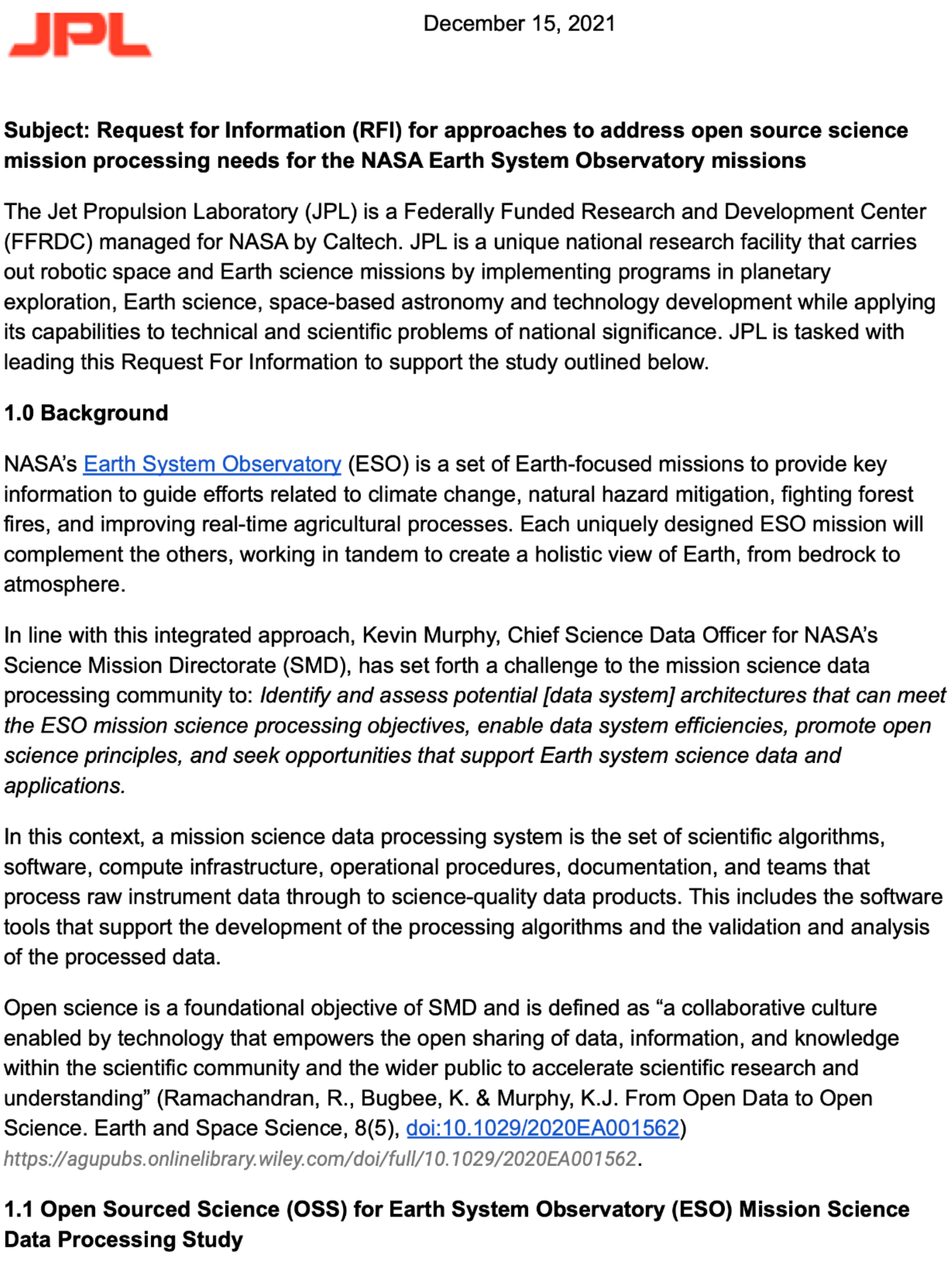
MODSIM	Modeling and Simulation
MMT	Metadata Management Tool
NASA	National Aeronautics and Space Administration
NCAP	NESDIS Cloud Archive Program
NCCF	NESDIS Common Cloud Framework
NCEI	National Centers for Environmental Information
NCIS	Cloud-sandbox Infrastructure Services
NESDIS	National Environmental Satellite, Data, and Information Service
NEX	NASA Earth Exchange
NGE	NESDIS Ground Enterprise
NISAR	NASA-ISRO Synthetic Aperture Radar (SAR)
NIST	National Institute of Standards and Technology
NOAA	National Oceanic and Atmospheric Administration
NOS	New Observing Systems
NPR	NASA Procedural Requirements
NRT	Near Real Time
NSF	National Science Foundation
OBPG	Ocean Biology Processing Group
OCI	Ocean Color Instrument
OCO	Orbiting Carbon Observatory
OGC	Open GeoSpatial Consortium
ORNL	Oak Ridge National Laboratory
OSI	Open Source Initiative
OSS	Open Source Science

OVF	Open Virtualization Format
PACE	Plankton, Aerosol, ocean Ecosystem
PAL	Product Algorithm Laboratory
PB	Petabytes
PEST	Policy, Economics, Sociocultural Factors, an Technologies/Tools
PGE	Program Generated Executables
PI	Principal Investigator
PiaB	Pipeline in a Box
PO.DAAC	Physical Oceanography DAAC
POR	Program of Record
PPM	Part Per Million
R&A	Research and Analysis
RFI	Request For Information
RGT	Reference Ground Track
ROSES	Research Opportunities in Space and Earth Sciences
RTC	Radiometric-Terrain Correction (SAR Data product)
S3	Simple Storage Service (associated with AWS)
SAR	Synthetic Aperture Radar
SAT	Science Activity Timeline
SAWG	System Architecture Working Group
SBG	Surface Biology and Geology
SMCE	Science Managed Cloud Environment
SDAP	Science Data Analytics Platform
SDC	Surface Deformation and Change

SDS	Science Data System
SDST	Science Data Support Team
SIPS	Science Investigator-led Data System
SLC	Single Look Complex (SAR data product)
SMD	Science Mission Directorate
SNPP	Suomi National Polar-orbiting Partnership
SOC	Science Operations Center
SPD	Science Mission Directorate Policy Document
SPS	Science Planning System
SQL	Structured Query Language
SQS	Simple Queue Service
ST	Science Teams
STSci	Space Telescope Science Institute
STAC	SpatioTemporal Asset Catalog
SWOT Analysis	Strength, Weakness, Opportunity, Threat
SWOT	Surface Water and Ocean Topography
TB	Terabyte
TCO	Total Cost of Ownership
TESS	Transiting Exoplanet Survey Satellite
TIR	Thermal Infrared
TOPS	Transform to OPen Science
TPU	Tensor Processing Units
TRISHNA	Thermal infraRed Imaging Satellite for High-resolution Natural resource Assessment

TRL	Technology Readiness Level
TROPICS	Time-Resolved Observations of Precipitation structure and storm Intensity with a Constellation of Smallsats
UKSA	United Kingdom Space Agency
USGS	United States Geological Survey
USML	United States Munition List
VIIRS	Visible Infrared Imaging Radiometer Suite
VPC	Virtual Private Cloud
VNIR	Visible and Near-Infrared
VSWIR	Visible to Short-Wave Infrared
V&V	Validation and Verification
WBS	Work Breakdown Structure
WPS	Web Processing Service
XML	eXtensible Markup Language
XSEDE	eXtreme Science and Engineering Discovery Environment
YARN	Yet Another Resource Negotiator
YOOS	Year of Open Science

Appendix - Request For Information (RFI)



JPL

December 15, 2021

Subject: Request for Information (RFI) for approaches to address open source science mission processing needs for the NASA Earth System Observatory missions

The Jet Propulsion Laboratory (JPL) is a Federally Funded Research and Development Center (FFRDC) managed for NASA by Caltech. JPL is a unique national research facility that carries out robotic space and Earth science missions by implementing programs in planetary exploration, Earth science, space-based astronomy and technology development while applying its capabilities to technical and scientific problems of national significance. JPL is tasked with leading this Request For Information to support the study outlined below.

1.0 Background

NASA's Earth System Observatory (ESO) is a set of Earth-focused missions to provide key information to guide efforts related to climate change, natural hazard mitigation, fighting forest fires, and improving real-time agricultural processes. Each uniquely designed ESO mission will complement the others, working in tandem to create a holistic view of Earth, from bedrock to atmosphere.

In line with this integrated approach, Kevin Murphy, Chief Science Data Officer for NASA's Science Mission Directorate (SMD), has set forth a challenge to the mission science data processing community to: *Identify and assess potential [data system] architectures that can meet the ESO mission science processing objectives, enable data system efficiencies, promote open science principles, and seek opportunities that support Earth system science data and applications.*

In this context, a mission science data processing system is the set of scientific algorithms, software, compute infrastructure, operational procedures, documentation, and teams that process raw instrument data through to science-quality data products. This includes the software tools that support the development of the processing algorithms and the validation and analysis of the processed data.

Open science is a foundational objective of SMD and is defined as "a collaborative culture enabled by technology that empowers the open sharing of data, information, and knowledge within the scientific community and the wider public to accelerate scientific research and understanding" (Ramachandran, R., Bugbee, K. & Murphy, K.J. From Open Data to Open Science. Earth and Space Science, 8(5), doi:10.1029/2020EA001562) *https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2020EA001562*.

1.1 Open Sourced Science (OSS) for Earth System Observatory (ESO) Mission Science Data Processing Study

In order to address the challenge laid out above, SMD has established a Study Team tasked to provide a recommendation on a mission science data processing system architecture for the ESO missions.

The Study Team has been assigned the following objectives:

1. Understand the data processing system needs and goals of the ESO Missions and NASA program offices (due: November, 2021).
2. Understand the current approaches and future trends in big-data processing systems and open science (due: April, 2022).
3. Identify and evaluate system architectures and implementation approaches (due: August, 2022)
4. Issue a mission science processing system architecture recommendation (due: September, 2022).

Objective #1 was completed through the successful execution of Workshop #1 on October 19-20, 2021. At this workshop the Study Team received input from NASA Program Offices and the ESO Missions regarding requirements, constraints, recommendations, and opportunities for science data processing. A copy of the presentations, a recording of the entire workshop, and a report issued by the Study Team is available on the Study website here: https://go.usa.gov/xe7GR.

The Study Team is now focused on addressing Objective #2 and are seeking input from organizations with relevant expertise in big-data processing and open science to help guide the study. The approach to supporting Objective 2 is to hold a second workshop (Workshop #2, on March 1-4, 2022) with invited participants presenting on key topic areas that will inform and advance approaches to developing and enabling open science mission processing systems. Invited participants will present on the topic areas outlined below, along with presentations from current and future NASA Earth mission science processing teams.

2.0 Scope

The Study Team is seeking input from commercial and non-commercial organizations with demonstrable expertise in processing scientific data or comparable datasets and/or implementing successful open science approaches.

Our goal is to broaden participation of historically excluded communities through encouraging responses from minority serving institutions, and small disadvantaged businesses to receive diverse inputs to this study. We encourage all responses to discuss how they serve these communities and any lessons learned.

The combination of ESO missions in the formulation stage along with the commitment to and foundational objective of advancing open science presents an enormous opportunity. This RFI calls for responses that may not only influence the approach to and delivery of mission science processing capabilities, but identify transformative approaches to how systems are designed and used to increase science and applications return. The study team and NASA leadership are looking forward to the responses to this RFI and the recommendations resulting from the overall study.

The OSS for ESO Mission Science Data Processing Study Team seeks input from a broad and diverse set of organizations on recommendations for approaches and capabilities that address the following topic areas:

1. **Data Processing System Architecture**. This refers to a software system architecture that supports processing of petabyte scale data in near real-time (hours) collected from a variety of data sources. Responses to this topic should provide an overview of an operational processing system, describe the components of the architecture, how the components interact, component and system dependencies, maturity for use in operations, and relevance to this Study.
2. **Open Science**. This refers to the technology-based methods and solutions that empower the open sharing of data, information, analytics, and knowledge within the scientific community and the wider public to accelerate scientific research and understanding. Responses to this topic should describe capabilities that a mission science data processing system should adopt in order to enable open science such as containerization for reproducible workflows, metadata curation, intellectual property permissions, version control, etc.
3. **Component Technologies.** This refers to transformative software and computing capabilities and methodologies that have the potential to increase efficiencies, promote interoperability, or enable new processing functionality. Responses to this topic should describe capabilities that have been developed and quantify how the capability might transform a mission science data processing system.
4. **Downstream Interoperability.** This refers to functions that a mission science data processing system should provide in order to support downstream consumers of the output data products. Responses should describe downstream data or information systems that will potentially utilize ESO data products in support of Earth system science and/or applications and how the system will likely interface with an ESO mission science data processing system. Additionally, responses should include recommendations on functions that the mission science data processing should provide in order to lower the barriers for use, integration, and analysis of data products.
5. **Other Recommendations**. This topic area allows for responders to provide recommendations on relevant mission science data processing topics not specified in Scope items 1-4 which might be considered critical to meet the goals of the study.

3.0 Responding to this RFI

Responses to this RFI should be submitted as follows:

- Send an e-mail to Glenn.E.Campbell@jpl.nasa.gov with the subject line: "OSSPS Workshop #2 RFI Response"
- In the body, include the public Data Object Identifier (DOI) that was prepared using the detailed instructions below

The submission must adhere to the following guidelines for content length:

1. Name, description and point of contact of the organization, relevant past experience and background information. 1 page.

2. A response to one or more topic areas outlined in Section 2.0, 1 page in length for each topic area being responded to (i.e. no more than 5 total pages if responding to all five topic areas). Please note the topic area(s) responded to on each submission.

3. Links to publicly available publications and relevant resources that support the responses. 1 page.

The requested information is for preliminary planning purposes only and does not constitute a commitment, implied or otherwise, that JPL will solicit you for such procurement in the future. Neither JPL nor the Government will be responsible for any costs incurred by the respondent in furnishing this information. Following in the spirit of SPD-41 (https://go.usa.gov/xeARr), prospective respondents are advised that any information provided shall be public.

Responses to this RFI are due by 11:59 pm (PST) on February 1st, 2022. In compliance with NASA OSS policy and to provide a fair and transparent study, responses must be submitted as a public DOI. Respondents may acquire a public DOI through their institution or https://zenodo.org/. Please add the following 'conference' details when you create your zenodo DOI. To get a Zenodo DOI sign in https://zenodo.org/ and click 'upload'. Upload your content as a .PDF, fill out details and at the bottom of the form please click on 'Conference' and fill out the following information: Conference title: Open Sourced Science (OSS) for Earth System Observatory (ESO) Mission Science Data Processing Study. Dates: 1-4 March 2022. Website: https://earthdata.nasa.gov/esds/open-science/oss-for-eso-workshops.

4.0 Disposition

The Study Team Steering Committee will review each response and evaluate it for inclusion in the upcoming Workshop #2, all are welcome to participate. Each response to a topic area will be assigned one of the following ratings after evaluation:

- Highly relevant. The response contains information that is highly aligned with the goals of the Study and warrants deeper consideration by the Study Team. Responders will be

invited to present their responses at Workshop #2 on March 1-4, 2022 and may participate in follow-on discussions.
- Relevant. The response contains information that supports the goals of the Study and will be considered by the Study Team. Responders may be contacted at a later date for more information.
- Out of Scope. The response does not contain information that supports the Study.

The review dispensation will be completed and the results published on the study Website. Invited presenters of Workshop #2 will be notified by February 15th, 2022, and participation is voluntary. The workshop will be held virtually to allow all presenters the ability to participate from their remote location and include the broader community in participation.

Sincerely,

Glenn E. Campbell

Glenn.E.Campbell@jpl.nasa.gov

Project Acquisition Manager, Acquisition Division

Office (818) 354-2530; Mobile: (818) 648-9764