The Value of Hosted JupyterHubs in enabling Open NASA Earth Science in the Cloud

Response to Aspect 1, the question on user needs and use cases for scientific data and computing in support of Open Science at SMD

Relevant NASA SMD scientific Division: *Earth Science*

Submitter:

Cassandra Nickles / Jet Propulsion Laboratory, California Institute of Technology / cassandra.l.nickles@jpl.nasa.gov

Authors (Name / Affiliation / email / ORCID):

Cassandra Nickles / Jet Propulsion Laboratory, California Institute of Technology / cassandra.l.nickles@jpl.nasa.gov / 0000-0001-9930-1433
Aaron Friesz / NASA Land Processes DAAC / afriesz@contractor.usgs.gov / 0000-0003-4096-3824
Alexis Hunzinger / NASA Goddard Earth Sciences Data and Informations Services Center / alexis.hunzinger@nasa.gov / 0000-0003-2369-5370
Andrew P. Barrett / National Snow and Ice Data Center, CIRES, University of Colorado at Boulder / andrew.barrett@colorado.edu / 0000-0003-4394-5445
Brianna Lind / NASA Land Processes DAAC / blind@contractor.usgs.gov / 0000-0002-5306-9963
Jess Welch / Oak Ridge National Laboratory / welchjn@ornl.gov / 0000-0002-5987-6387
Luis Lopez / National Snow and Ice Data Center / luis.lopez@nsidc.org
Mahsa Jami / NASA Land Processes DAAC / mjami@contractor.usgs.gov / 0000-0002-3594-3004
Michele Thornton / Oak Ridge National Laboratory / thorntonmm@ornl.gov / 0000-0002-6533-6328
Erin Robinson / Metadata Game Changers LLC, NASA Openscapes co-lead / erin@metadatagamechangers.com / 0000-0001-9998-0114
Julia Stewart Lowndes / Openscapes, NASA Openscapes co-lead / julia@openscapes.org / 0000-0003-1682-3872

NASA Openscapes Mentor Community

The following RFI presents collective thoughts of the NASA Openscapes mentor community. NASA Openscapes, co-facilitated by Erin Robinson and Julia Stewart Lowndes, is a space where members from seven NASA Distributed Active Archive Centers (DAACs) come together to collaborate. This cross-DAAC mentor community develops resources, teaches, and listens to feedback from NASA Earthdata end users on a regular basis. NASA has provided data freely through the DAACs for decades, enabling researchers to make significant contributions to understanding our planet that would not have been possible otherwise. As Earthdata migrates to the cloud, we at the DAACs have been positioned to help facilitate data ease of access for end users within this cloud infrastructure. We are uniquely familiar with the broader challenges our data end users face as they transition their workflows to the cloud and hope to share our

experiences and recommendations as NASA moves to inclusively support open science practices in the cloud framework.

The Problem: Cloud Transition Burdening the User Experience

NASA, along with many other federal agencies, is moving data to the cloud to enable analysis alongside the data. More and more, data is archived through a DAAC, not at a DAAC. Migrating NASA Earthdata to the cloud solves computing issues Earth system scientists frequently face regarding the access and processing of large/complex volumes and disparate data necessary for their research. For example, working with Earthdata on a local computer can lead to long (or impractical) download times and often require large storage capacities and enhanced processing systems. More storage and better computers yield proficient and faster science, potentially reinforcing societal inequities in science as not all researchers have access to the same institutional support or financial resources required. Working alongside the data in the cloud fundamentally solves some of these issues, but others persist.

While advances in cloud computing are exciting, in practice, individual users face a steep learning curve and still face inequities. Cloud workflows require new software tools and the skills, mindsets, and support to go with them. In addition, the constantly evolving nature of cloud infrastructure makes it challenging to get started without substantial support from experts. Often users must pay to have access to cloud computing platforms and/or align with institutions that have already set up shared cloud computing spaces. We see a future with cloud deployments. However, there are challenges listed as follows:

- Technology Gap: a growing gap between the technological sophistication of industry solutions (high) and scientific software (low). **this bullet is credited to Dr. Ryan Abernathy from Pangeo
- Resource Gap: Though NASA data is free and open to all, the transition to standing up consistent cloud computing platforms is not.
- Skills Gap: a growing gap between technical skills required to use the cloud and skills taught to researchers during training.
- Knowledge Gap: There are substantial changes to how data is accessed, processed, queried, stored, manipulated, and synthesized. Further, the file types, data structure, organization, and terminology are often novel.

After running several workshops, it became clear that there is a strong need (and desire from users) for NASA to facilitate access to cloud computing platforms and provide easy-to-use guides to help navigate the new environment. We should not need to be cloud infrastructure engineers to work with the data located in an S3 bucket in the Amazon Web Services (AWS) infrastructure, and users should have a cloud platform in which they can discover if using the cloud environment works for them without large investment. The real power of open science in the age of cloud computing is unleashed only if cloud platforms are accessible and accompanied by easy-to-use workflows that enable inclusive, efficient, and reproducible science. Unfortunately, as it stands today, scientists spend more time on the technicalities of the cloud rather than focusing on their important science.

Our Solution: The 2i2c Openscapes JupyterHub

A key objective of NASA Openscapes is to minimize "the time to science" for researchers. Cloud infrastructure can facilitate shortening this time. However, for many researchers, it can be a long leap from downloading and analysis on a local machine, to working in a cloud environment, particularly since many researchers across disciplines see coding skills as a large unmet need (e.g. Lowndes et al. 2017, Barone et al. 2017). To shorten this leap to a hop, we use a 2i2c-managed JupyterHub, which lets us work in the cloud next to NASA Earthdata in US-West-2.

2i2c is a nonprofit that designs, develops, and operates JupyterHubs in the cloud for research and education, including NASA Openscapes. 2i2c ensures that Hubs are cloud-vendor agnostic and are built using open-source software such as JupyterHub and Kubernetes. 2i2c also gives users the right to replicate their infrastructure. So, while our Openscapes JupyterHub is built on top of AWS, it could be replicated on GEE or Microsoft Azure, or ported to another AWS region. 2i2c also makes managing users easy with GitHub authentication. On our end, it takes less than 1 second to copy a user's GitHub username (or a list of names) into an approved list on the 2i2c Openscapes Hub for access. In contrast, in Spring 2022 when we experimented with the Science Managed Cloud Environment (SMCE), an AWS-based infrastructure for NASA-funded projects, the SMCE took about 4 minutes to add a single user. We also installed a GitHub GUI Addon within our JupyterHub that is helpful for learners and researchers as they learn how to set up work on the Cloud.

With this setup, we have flexibility to support a diverse range of user needs. The 2i2c Openscapes Hub has been used by us internally as a testing ground for developing cloud tutorials and workflows, but also externally in the research community for workshops like those for science teams and "Hackathons," a term used here to describe multi-day events with split time for teaching and helping researchers implement concepts into their research projects. The only software requirement to deploy the Hub is access to a computer and the internet.

Our JupyterHub is used in several categories:

- **We,** as a NASA Openscapes Mentor Community, use it to **build skills and comfort levels** with operating in the Cloud ourselves to **develop** Cloud tutorials
- **End-users**, including academic, government, non-profit, and industry researchers, use it to learn by following tutorials during **external training events**, for example, the 2021 Cloud Hackathon, 2022 ECOSTRESS Workshops, and 2022 Openscapes Champions Cohort for research teams.
- **DAAC Staff** use it to learn by following tutorials in **internal training events**, for example, at GES DISC, ESDIS SAFe train development, and for DAAC user working groups (UWGs).

All users have continued to have access to experimenting following these events; to date, we have not removed anyone from our JupyterHub. This last category of internal training for DAAC Staff has emerged as a new benefit of the Hub; without the 2i2c JupyterHub, many staff members have not yet been hands-on in the Cloud, and it is difficult to support others without that experience (slides from Hunzinger 2022, who was invited to share experiences at the LAADS DAAC: Early lessons learned from supporting end users' transition to the cloud).

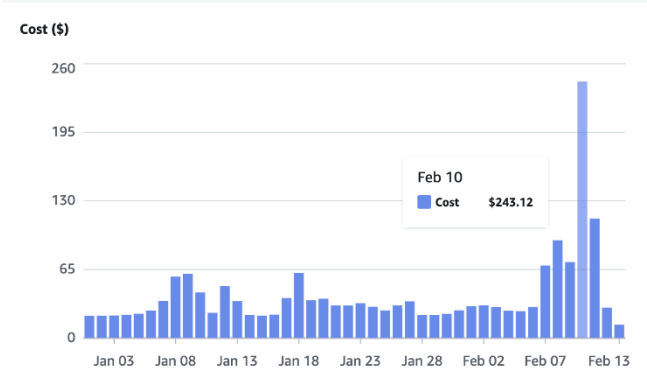

Figure 1. Example costs of the Openscapes Hub operations from January through mid-February 2023. EMIT Workshop hosted on February 10th.

For operations and maintenance costs, we have paid $10,000 per year to 2i2c directly as the base cost and we paid AWS an additional $10,000 in cloud credits. For a direct example, we recently hosted a workshop for EMIT data using our 2i2c JupyterHub. Leading up to the workshop, we would spend ~$20/day to run our instances among consistent users, while on the workshop day, we spent $243 to run 91 large instances (Figure 1). In addition to these costs, from a sustainability standpoint, it is necessary to have a dev ops visionary to manage the Hub, complementary to the support 2i2c provides. For us, Luis Lopez (NSIDC DAAC) has provided invaluable open science leadership and technical support working on our Hub.

Further, we have built in power and flexibility with the computing environment within 2i2c. Luis Lopez has developed corn, a base image that allows the provisioning of a multi-kernel Docker base image for Jupyterhub deployments. corn uses Pangeo's base image (a collection of scientific python packages widely used by the Earth Science community), installs all the environments it finds under ci/environments, and makes the environments available as kernels in the base image so users can select which kernel to use depending on their needs. As we worked with users, we found many Earth scientists have legacy code developed in other coding languages outside of Python and R, like Matlab. The different kernels within the JupyterHub cloud deployments are associated with these three common languages (Python, R and Matlab) to help facilitate workflow transfer. We are able to update this environment leveraging GitHub Actions and deployment as shown in Figure 2.

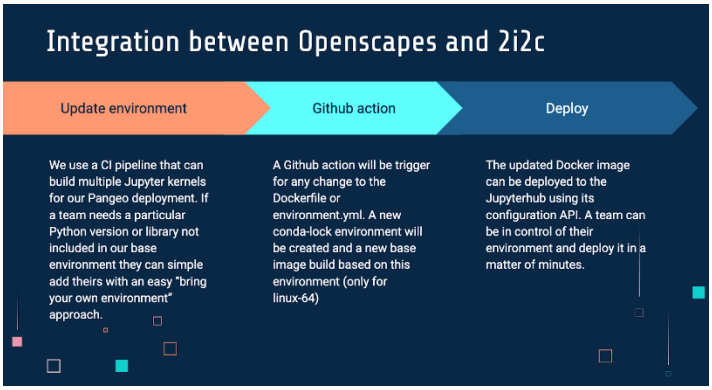

Figure 2. Integration between Openscapes and 2i2c. We update the environment via GitHub action and Docker deployment.

We acknowledge that using the Hub does not solve all problems, and as the NASA Openscapes community, we have also focused on addressing pain points for using the Hub identified by users and Mentors through all of this work. It is not always enough to get researchers within the cloud environment, we must also think about ways to make the data access experience within the cloud more accessible, decreasing the Skills and Knowledge Gaps expressed above. One of the participants of our 2021 Cloud Hackathon concluded we need "better documentation/tutorials for how to access data over the cloud. It would have been extremely difficult to do any of this

without the help of the hackathon." Collectively, we have responded by developing a number of resources, but there is still much more that needs to be done. We've made conceptual solutions that visualize workflows through Cheatsheets (Catalina Taglialatela and Cassandra Nickles, PO.DAAC). We also created a software solution: `earthaccess`, a python library developed by Luis Lopez that aims to simplify data discovery and access. This library reduces the need to know the intricacies of NASA's Application Programming Interfaces (APIs) and cloud data storage systems.

Conclusions & Recommendations

Shared cloud-hosted computational environments such as the 2i2c Openscapes JupyterHub allow multiple users to access cloud computing resources using tools and interfaces familiar to many researchers, decreasing the Technology and Resource Gap issues identified. These Hubs have been invaluable for NASA Openscapes, Mentors, staff members across the NASA DAACs, and researchers accessing NASA data in the cloud. Hubs are centrally-administered, removing the need for scientists to be cloud experts, reducing the barrier to entry, and allowing hundreds of users to access and analyze NASA data in the cloud for the first time. In addition to supporting end users, these Hubs provide a space for NASA staff to learn and develop resources to support researchers across the DAACs. Reproducibility also increases within shared cloud environments! Coding languages often require downloading specific packages to enable workflows, which may or may not have certain dependencies, etc., that can send users on rabbit trails lasting hours, time better spent on science and applications. Within the 2i2c environment, common packages have already been installed, removing barriers of potential frustration and decreasing time to science.

Within the last two years, we have the following main lessons learned from our end users:

- **The cloud learning curve is steep!** No one left our hackathons an expert, and we as DAAC staff continue learning and experimenting with this technology. We must do a better job laying a foundation with cloud basics and terminologies.
- **We need to provide resources that are easy to revisit.** A permanent and accessible cloud computing environment with learning materials would be highly utilized.
- **Continued support and education are critical**. It is necessary to host refresher workshops and even introduce better tools/methods that are rapidly developing.

In short, we need infrastructure like the 2i2c JupyterHub permanently to better support our Earthdata users. If the Earthdata Cloud is to be a key component of the ESDS Transform to Open Science (TOPS) program in this Year of Open Science and beyond, we must increase data accessibility. We recognize that costs for cloud computing infrastructure are real, and the operational complexity for the maintainers is not trivial. However, if we want Earthdata to be used in the cloud effectively, **we recommend that NASA recognize easy, accessible, and inclusive cloud access as a core service.** Our 2i2c Jupyter Hub has been critical for reducing barriers to cloud entry and having a shared environment to meet users where they are (see this blog by Luis Lopez for more). Could this successful model of a shared cloud computing platform be expanded in a broader way for all NASA Earthdata Cloud users? We must continue to close the loop between the users we work with and our engineers to build equitable solutions together.

References

Barone, L., Williams, J. & Micklos, D. (2017). Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. *PLOS Computational Biology* 13(11): e1005858. https://doi.org/10.1371/journal.pcbi.1005858

Lowndes, J., Best, B., Scarborough, C. Afflerbach J., Frazier, M. O'Hara, C., Jiang, N., Halpern B. (2017). Our path to better science in less time using open data science tools. *Nature Ecology & Evolution* **1**, 0160. https://doi.org/10.1038/s41559-017-0160