

ADVANCING OPEN SCIENCE THROUGH INNOVATIVE DATA SYSTEM SOLUTIONS: THE JOINT ESA-NASA MULTI-MISSION ALGORITHM AND ANALYSIS PLATFORM (MAAP)’S DATA ECOSYSTEM

Kaylin Bugbee¹, Rahul Ramachandran², Manil Maskey², Aimee Barciauskas³, Aaron Kaulfus¹, Dai-Hai Ton That¹, Katrina Virts¹, Kel Markert¹, Chris Lynnes⁴

¹University of Alabama in Huntsville

²NASA Marshall Space Flight Center

³DevelopmentSeed

⁴NASA Goddard Space Flight Center

ABSTRACT

Collaborative open science practices are changing the way research is conducted. These changes affect how scientists work together on data, code and information. Data systems enhance open science by offering forward thinking technological solutions, such as providing data and computation on the cloud, to enable collaboration, sharing and analysis. In this paper, we present our vision for a conceptual data system on the cloud that enables open science. We also present our work on the Multi-Mission Algorithm and Analysis Platform (MAAP) which has served as a pathfinder data system for this conceptual approach.

Index Terms— Data stewardship, cloud technologies, data systems, open science, open data

1. INTRODUCTION

Open science, or “the idea that scientific knowledge of all kinds should be openly shared as early as it is practical in the discovery process,” [1] has been a topic of much discussion within the scientific community of late [2]. While the scientific community has always attempted to maintain some level of openness through the journal publication process, open science as a topic has gained more attention due to the Science 2.0 phenomenon where scientific information and data can be communicated via the Internet [2]. These technological advances have changed the way scientists work and collaborate. Through the Internet, scientists work together by quickly sharing data, code and information. This ease of sharing has ensured that scientific research is no longer confined to a lone researcher in a lab and is instead becoming much more cooperative.

One scientific community that wants to work in a collaborative and open science environment is the global aboveground terrestrial carbon dynamics, or biomass,

community. This community is committed to understanding the distribution in and changes to carbon on the Earth’s surface. Space-borne Earth observation data, in combination with field and airborne observations, are improving the understanding of these processes by providing improved geographic coverage, better spatial resolutions and an ongoing period of temporal observations [3]. However, there are some barriers to use of these key data for the research community. For example, not all relevant data is openly shared and the relevant data which is open is dispersed across various archives and organizations, making integration challenging and time consuming. Integration is further complicated by the heterogeneous nature of the biomass community’s data. Data is collected at multiple spatial scales and different geometries and is provided in various formats. Once data is collected and integrated, collaborating on biomass algorithms and the output data resulting from these algorithms is cumbersome.

Data systems have an opportunity to enable the open data, information and compute needs of the biomass community and other research communities by providing forward thinking technological solutions that lower barriers to research. One such solution includes moving data from on-premises storage to the cloud. Placing data on the cloud provides a number of new capabilities which may address a number of open science research needs. However, moving away from on-premises data systems requires a rethinking of both the conceptual architecture for a data system and the supporting data stewardship activities of the system. In this paper, we present our conceptual architecture of a data system on the cloud. The Multi-Mission Algorithm and Analysis Platform (MAAP) is presented as a pathfinder data system for this conceptual approach. Lessons learned from the MAAP development are shared along with future work for the project.

2. CLOUD-BASED CONCEPTUAL DATA SYSTEM ARCHITECTURE

A cloud-based data system includes core traditional data system components but also other new components and capabilities that are now possible due to the cloud environment (Fig. 1). The traditional data system components, shown in gray in Figure 1, include a data lake and associated core data services. A data lake is defined as a large repository of heterogeneous raw data that serves as a central repository where data of all formats and from various external sources are stored without a strict schema [4]. Core data services include essential data stewardship activities such as providing a metadata catalog, ensuring access to supporting data documentation and offering data visualization options. These core services lower barriers to data use by ensuring data is discoverable, accessible and usable. In addition to the data lake and the core services, the MAAP also provides client access via application programming interfaces (APIs) and tools. Client access serves as a gateway between the various other components.

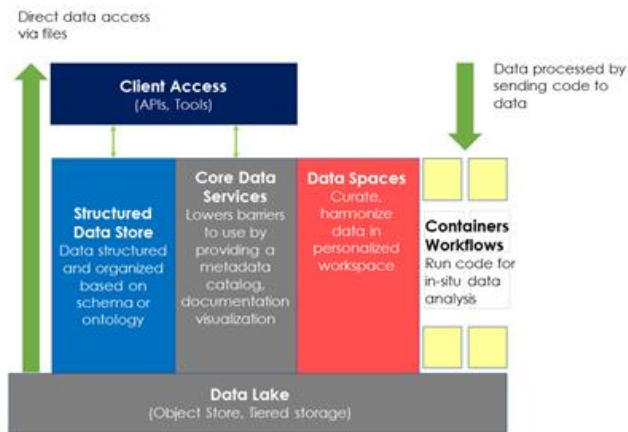


Fig. 1: The conceptual data system architecture in the cloud.

The cloud-based data system also includes new components such as data spaces (shown in red in figure 1), containers and workflows (shown as yellow boxes in figure 1) and structured data stores (shown in light blue in figure 1). A data space is a private personalized workspace where users can curate and integrate key data for a given research question. Data in a data space may originate from the data lake or may be uploaded from a user's private machine. Data spaces are highly coupled to containers and workflows. Data analysis is conducted by sending the code to the data in situ. This coupling minimizes direct file download by enabling analysis within the platform and also makes analysis at scale possible. The results of analysis conducted using containers and workflows may also reside in a data space. Last, the data system also includes structured data stores. As opposed to the data lake, the structured data store provides more order, and therefore optimization, to the data by organizing the

data around a schema or ontology. The structured data store allows users to more efficiently analyze large amounts data in order to gain more interactive insights [5]. We envision that future data systems will be combinations of these capabilities that can be layered together. Research community needs and technological implementation choices will drive the selection of the components.

3. CLOUD-BASED DATA SYSTEM STEWARDSHIP ACTIVITIES

Data stewardship needs for a cloud-based, collaborative, open science data system share some similarities with a traditional data system but there are also some distinct differences. Traditionally, data systems and archives have been required to write and maintain data management plans (DMPs). A DMP addresses the management of Earth science data from the time of their collection to their entry into an archive [6]. A DMP includes information on data access, quality, preservation and data stewardship activities such as metadata, documentation and persistent identifier assignment. While all of the traditional DMP information is essential, it is often limited to data governance within only the data lake and core data services. Governance, however, is needed for all components of a cloud-based data system including the data spaces and the structured data stores. This broader concept of data governance should also include tracking the lineage of data especially as it moves from data lakes to structured data stores and other services. Therefore, we envision a data governance plan that encompasses all components in the system not just the core traditional data system components.

The scope of stewardship activities also needs to be expanded for open science data systems. Traditionally, data systems have primarily focused on providing robust stewardship services for data while software, code and algorithms have provided a supporting role to the data. However, scientists are increasingly collaborating on code and algorithms and cloud-based data systems make it easier to run an analysis on demand as opposed to simply storing final data outputs. Stewardship activities for software, code and algorithms will need to be more robust to allow discovery and reuse of these key research objects. These stewardship activities include robust metadata, documentation and preservation plans for software, code and algorithms.

4. MAAP DATA SYSTEM IMPLEMENTATION

The Joint ESA-NASA Multi-Mission Algorithm and Analysis Platform (MAAP) is a collaborative, cloud-based, open science platform dedicated to the unique research

needs of the biomass community. The MAAP also represents a new approach to accessing, sharing and processing data. The platform is jointly managed by ESA and NASA with each agency developing separate but connected systems. The NASA MAAP cloud-based data system implementation is presented below.

4.1. MAAP Data Lake and Core Data Services

Data is ingested into the NASA MAAP's data lake from two primary sources. The first of these sources are trusted public data archives which distribute relevant data that is not currently available on the cloud. These data are ingested into the data lake by the NASA MAAP data team using the methodology described in [7]. The second source of data for the data lake is users' data spaces. MAAP users may elect to share data to the broader platform for discovery and reuse. To support this type of sharing, data is ingested from a user's data space into the NASA MAAP data lake using the same ingest framework that is leveraged for the archived data. While both the user data and the archived data are required to comply with some data format requirements, the various data collection methods and sensors make the NASA MAAP data lake heterogeneous in nature.

For the MAAP, data in the data lake is discoverable via a centralized metadata catalog. The centralized catalog includes metadata describing data objects within NASA's data lake as well as within ESA's data lake. Metadata is also created for user shared data and is included in the centralized catalog. This centralized metadata catalog ensures discovery and open access to all relevant biomass data regardless of organizational source.

Other core data services provided by the NASA MAAP data team include links to relevant data documentation and fast browse visualization capabilities. To date, prototype fast browse visualization capabilities have focused on developing solutions to quickly visualize large data volumes with minimal information loss and efficient performance.

4.2. MAAP Data Spaces and Containers/Workflows

For the NASA MAAP, a private Eclipse Che [8] environment serves as a user's data space. Within this environment, the user can upload data from a local machine, integrate data from the data lake using the MAAP API or leverage containers and workflows to run analysis at scale. The MAAP's containers and workflows, in combination with the data lake, allow scientists to upscale algorithms globally, which meets another major science goal for the platform. Also, data within the data space may be privately shared with collaborators or, as previously mentioned, may be more broadly shared to the MAAP user community. This capability meets another major scientific goal of making collaboration easier.

4.3. MAAP Structured Data Store

For the NASA MAAP, the data team prototyped a small PostgreSQL database on the cloud which contains biomass relevant calibration and validation data. The MAAP API enables efficient data access to the database via queries from within a user's data space. Longer term, analysis optimized data services and other solutions for big data analysis will be developed for the MAAP to provide improved and efficient performance for users.

4.4. MAAP Data Stewardship

Recognizing the need to document data governance beyond the data lake and core data services, the MAAP team is drafting a data governance plan for the data system. The governance plan includes many of the traditional DMP components for the data lake and the core data services but also documents other key data policies for the other system components. The documentation includes information on who can access data, the quality of the data, how the data is stored, how the data can be accessed and the lineage of the data. The MAAP data governance plan will continually evolve as technological and user needs change. Lessons learned from drafting and maintaining the plan will inform the broader community as more data systems move to the cloud.

4. FUTURE WORK

The MAAP is currently being tested and used by a select group of biomass subject matter experts. Development of the MAAP data system components and capabilities will continue as the platform moves towards a fully operational system. An emphasis will be placed on further prototyping and developing structured data store solutions that meet the research community's needs. The team will also operationally implement the data governance plan in order to ensure the MAAP is adhering to good data stewardship practices that are appropriate for a cloud-based system.

The data team is also planning to expand the core data services provided by the MAAP to include discoverability of biomass relevant software and documentation in addition to data. The team will be drafting a metadata model to support the discovery of these objects with a longer term goal of establishing linkages between data, software and documentation. This new core stewardship activity is essential to making open science collaboration and sharing of software and code easier.

Longer term, the MAAP data team envisions completing technology transfers with other projects and collaborators in different scientific domains [9]. We can envision other platforms like the MAAP being created with a different scientific priority and which addresses the unique data and

analysis needs associated with that community. These other platforms may use any or all of the data system components as described above to meet a given need.

5. CONCLUSIONS AND LESSONS LEARNED

Our conceptual cloud-based data system architecture includes a singular data lake. However, as more data systems move data to the cloud, the data lake may evolve to become a federated data lake depending on the needs of the data system. For example, the MAAP utilizes a single data lake that requires ingestion of data from various sources into the MAAP data lake. This practice is required since most of the data needed by the biomass community is not available on the cloud. While a single data lake streamlines discovery and use of data in a cloud-based environment, it is less than ideal from a data stewardship perspective since it requires that data be monitored for new versions and updates. Longer term, the MAAP data lake will become a federated data lake to ensure data timeliness. However, moving from a singular data lake to a federated data lake will most likely pose unique challenges from a cost perspective.

Enabling discovery and use of structured data stores within a data system requires some consideration. For example, in the MAAP, a link to the prototype Postgres database was provided in the collection level metadata of the associated data product. This solution is less than ideal since it requires a user to first find the data product and then find the access link to the structured data store buried within the metadata. While it is appropriate to provide a link to services within collection level metadata, a longer term solution will allow users to quickly discover structured data stores independently of the data.

Cloud-based open science platforms, as in the case of the MAAP, may have new sources of data such as user shared or citizen science data. Balancing stewardship and service needs for these types of data with the effort required by the user to meet those needs is a delicate task. The community wants to share and collaborate on data quickly and easily. For example, the NASA MAAP data team has lowered the effort required for sharing data by streamlining required metadata fields and by providing automated file level metadata extraction. Data services for user data, such as fast browse capabilities, have not been developed to date but are desired by the community. To ensure a baseline level of service and stewardship, cloud-based open science platforms should consider developing automated workflows and easy to understand requirements for users.

The global aboveground terrestrial carbon dynamics research community is just one of many scientific communities that want to work in new and collaborative ways. The MAAP represents a first step in understanding how data systems can more broadly support open science

activities and how cloud-based data systems need to evolve to support these activities.

6. REFERENCES

- [1] Scholarly Communication: Open Research. University of Cambridge. <https://osc.cam.ac.uk/open-research>. Accessed 11/20/19
- [2] S. Bartling and S. Friesike. (2014) "Towards Another Scientific Revolution." In: *Opening Science*. Springer, Cham. https://doi.org/10.1007/978-3-319-00026-8_1
- [3] Herold, M., Carter, S., Avitabile, V. et al. *Surv Geophys* (2019) 40: 757. <https://doi.org/10.1007/s10712-019-09510-6>
- [4] P. Sawadogo, E. Scholly, C. Favre, E. Ferey, S. Loudcher and J. Darmont. "Metadata Systems for Data Lakes: Models and Features." *1st International Workshop on BI and Big Data Applications*, Slovenia, 2019, pp.440-451. doi: 10.1007/978-3-030-30278-8
- [5] R. Ramachandran, K. Bugbee, M. Maskey and C. Lynnes. "From ARD to AODS: Future of Analytics for Earth Observations," *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, Yokohama, Japan, 2019, pp. 4783 – 4786.
- [6] Data Management Plan Guidance. NASA Earthdata. <https://earthdata.nasa.gov/collaborate/new-missions/data-management-plan-guidance>. Accessed 1/13/2020.
- [7] K. Bugbee *et al.*, "Building a Data Ecosystem: A New Data Stewardship Paradigm for the Multi-Mission Algorithm and Analysis Platform (MAAP)," *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, Yokohama, Japan, 2019, pp. 4261-4264. doi: 10.1109/IGARSS.2019.8898070
- [8] How Eclipse Che Works. <https://www.eclipse.org/che/technology/>. Accessed 01/14/2020
- [9] P. Edwards, S. Jackson, G. Bowker and C. Knobel. "Understanding Infrastructure: Dynamics, Tensions, and Design," (2007).