Why Use the Cloud? An Earthdata Vision

Introduction

Large remote sensing missions, archived by the EOSDIS (Earth Observing System Data and Information System) DAACs (Distributed Active Archive Center), are challenging the end user's traditional download and access paradigm. While the core functionality of the DAACs (ingest and archive, cataloging, and data access) will always be available, the Earth Science Data and Information System (ESDIS) project office is also exploring new ways to enable science and the transformation of data into knowledge and information. The cloud offers a scalable and effective way to address storage, network, and data movement concerns while offering a tremendous amount of flexibility to the user. The new big missions, Surface Water and Ocean Topography (SWOT) and NASA-ISRO SAR (NISAR), will push EOSDIS into its next phase — co-located, multidisciplinary, cloud based archive and distribution centers that enable analysis next to the data. Large multiyear global datasets from other ongoing missions will also become available in cloud environments. These will be scheduled for availability in the cloud when use cases suggest research will benefit or be enabled. Users should look to the DAACs for information about which datasets are currently hosted in the cloud and which datasets are targeted for hosting in the cloud environment.

NASA EOSDIS is evolving to use a commercial cloud to ingest, archive, process, distribute, and manage the anticipated large volumes of new mission data. Placing the EOSDIS archive collectively in the cloud will, for the first time, place NASA Earth Observing data "close to compute" and improve management and accessibility of these data while also expediting science discovery for data users. Having EOSDIS data in the cloud will not change the existing user experience of interacting with these data, but it will offer new methods of access not otherwise possible with on-premises platforms.

It is important to note that under NASA's full and open data policy, all NASA mission data (along with the algorithms, metadata, and documentation associated with these data) must be **freely available and provided to the**

Cloud technology is evolving so fast that it is likely that some details in the primer may no longer match reality when you are trying to use it. If you find mismatches (e.g. broken third-party links), please send them to support@earthdata.gov so that we can feed them into the next release of the primer.

public as soon as possible following a checkout period to ensure data accuracy and validity. As such, **all NASA data is free to access and download, regardless of whether the user is working on the cloud or not**.

Users have a range of cloud computing platforms to choose from if they wish to perform some or all of their science or applications workflows in a cloud-based environment. NASA's Office of the Chief Information Officer selected Amazon Web Services (AWS) as the source of general-purpose cloud services for NASA, including ingest, archive, processing, distribution, and management of data. In addition, NASA EOSDIS is also working with Google Earth Engine (GEE) to make NASA data accessible in the GEE cloud-based analysis platform.

New Missions, Big Data

Big Data is the largest reason, pun intended, for the move to the cloud. Upcoming missions, such as SWOT and NISAR shift the scales of data archived at the DAAC. The SWOT mission will produce over 60 PB of data over the life of the three-year mission, while NISAR will generate 140 PB of data over its three-year scheduled mission. These are orders of magnitude larger than anything users have seen in the remote sensing world, and represent a challenge for the DAACs, EOSDIS, and the end user community. Let's not forget extended missions, often the norm, need to be addressed. If valuable science products are coming down the pipeline, a lack of resources on the DAAC side should not be the reason a mission does not continue onward.

Aside from these massive missions, we are seeing near term missions, such as new additions to the Sentinel series, breach the 1 PB mark. Storage, network and data movement capabilities are easily taken for granted until the size of the data becomes unwieldy.

Current missions, such as Terra, have observation products which span over 20 years. Their data provide an opportunity to research geophysical trends, discover events, and find phenomena in relatively long time series studies. Having the entire mission collection in a cloud environment offers researchers advantages such as being able to effectively scale their resources to perform relatively quick data analyses covering the entire globe, over the entire mission collection. This is currently not possible with the current user download paradigm.

Storage

In order to do long time series data analysis, users typically download a copy of all of the mission data onto their local storage facility. If their research is looking at observations from multiple instruments and multiple missions, then local storage is required for each. This allows them to run research algorithms multiple times over the same data, iteratively, until they are satisfied with the algorithm and results. If one or more of the collections are available in the cloud environment, the user can avoid local storage resources and costs by accessing the data directly from their cloud compute accounts - as long as their account is within the same cloud environment as data (e.g., user's cloud compute account is with Amazon Web Services (AWS) and in the same AWS region as where DAAC data is stored). For more details regarding the cost model in the new cloud paradigm, see ***Understanding and Managing Costs in the AWS Cloud*** tutorial.

Network

In the current user download paradigm, access is limited by the current bandwidth provided between the DAAC and Internet and between the Internet and user's facility. When DAAC data is hosted in the cloud environment, and users have compute accounts in the same cloud environment, users can take advantage of the high bandwidths available within the cloud environment. These are typically much higher than what is currently supported by DAAC-to-Internet and Internet-to-user. This is what allows users to obtain rapid access to large volumes of mission data.

Much like storage, network capacity is a critical component for the Earthdata enterprise, both for the distribution of data to the end users and for the ingest of the science data products. The cloud offers the ability to scale our network with demand. For the NISAR and SWOT missions, the data products will be produced within the same cloud the DAACs are running in, and therefore users will have access and network bandwidth at the fastest possible speeds offered by the cloud provider (in this case AWS).

Compute Resources

In the current paradigm, users' processing resources have a prescribed capacity that ultimately determines the time it will take to perform data analysis. Facility refresh is necessary on a regular basis to maintain state-of-the-art speeds. Between studies, while algorithms are being refined or created, a user's resources may be idle - but still need to be administered to and maintained. In shared environments, users may need to compete for processing time, and can be subject to higher priority activities. In the cloud environment, compute resources can be requested and provisioned on an as-needed basis and once again, we refer to the user generically working within the general AWS. Once the user's research algorithms have been tested and are ready for production, cloud resources can be provisioned to scale up. Once the research results have been obtained over the entire mission, the resources can be released. This pay-as-you-go enables maximum flexibility in resource utilization, and permits scaling beyond the capacity of a user's facility to get results faster without committing to the long-term cost of upgrades and maintenance of an in-house facility.

Data Movement and Pre-Processing

Combined with the storage and network limitations discussed above, the ESDIS project office selected a cloud solution to best enable end user science, whether that happens at a university research lab, an international non-profit, a startup looking to create innovative solutions through value-added data products and analysis, or a graduate student working on their thesis. While the cloud may not be ideal for everyone (indeed there are economic and technological situations where users may not want to use the cloud at all, and we will address them later), it does offer solutions to a problem that these large NASA missions, and the need for heterogeneous or multi-source data, present.

By moving their processing to the cloud, users could address several of their challenges when handling Big Data:

1. The user with limited bandwidth can find, access, and process all of the data required without need to download data to their local computers*. Only resulting data products or the final calculations and summaries need to be downloaded. This represents the download of information or knowledge as opposed to the download of data.
2. Users without the resources to download and store data can use the cloud to do their processing without having to pay upfront for storage or computer hardware. Computer can be spun up and down to minimize cost.
3. Users who can afford the upfront costs may still opt to use the cloud if data is not needed in perpetuity, or if massive amounts of computational power is required over a small amount of time. This is similar to the economic model of reserving time on a high performance computing (HPC) platform, though architecturally there are major differences.

*Rest assured that these cloud-based DAAC services for finding, processing, and accessing the data required for their needs can also be used in the case where the users wish to download the data and perform the science or application on a local machine or cluster because that best suits their needs.

To address data movement issues, the DAACs will provide services to minimize the amount of data moved by the users. These services will be responsive to data transformations required for the scientific analysis and utilization of data. This addresses the "80/20 rule" [1], whereby 80% of research or data analysis is spent acquiring and cleaning it for a specific use, by flipping the rule on its head. In other words, these DAAC-provided, cloud-based services that are co-located with the data aim to ease the pre-processing, transformation, and treating of the data for a specific use, so that the user can spend the 80% focusing on the science or application work instead. To provide these services, the DAACs themselves will leverage the compute power offered by the cloud - as demand for a service or dataset spikes (e.g. a flood event or hurricane) more services can be deployed to handle the demands of the user.

Increase Science Impact

As shown above, there are a multitude of technical reasons for moving to the cloud. Non-technical reasons focus mainly on increasing the science value of DAAC data holdings by utilizing cloud computing.

Multidisciplinary Data Access

Currently, the DAACs housing NASA data are spread across the country (see Figure 1 below). While the data can be stored in isolation, end users require multiple data products to inform their research and decision making. Different DAACs use different interfaces and access mechanisms, requiring users to understand the nuances of each DAAC in order to retrieve data for their use. This is another example of work going into the 80% of the 80/20 rule. While this in the past was simply a hassle, with the advent of multi-petabyte data, this becomes a substantial roadblock.

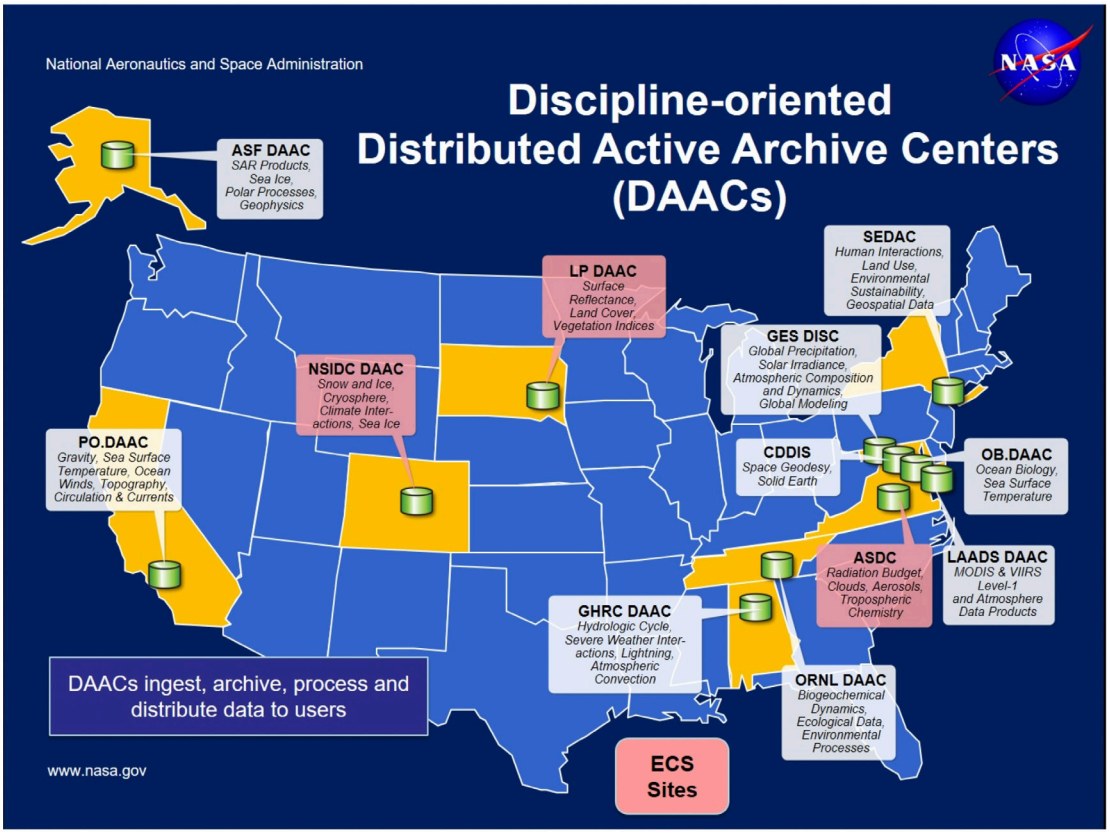

Figure 1: Map of DAAC Locations

For instance, if the Alaska Satellite Facility (the DAAC hosting NISAR data) and the PO.DAAC (the DAAC hosting SWOT data) house their data on-premises in Fairbanks, AK and Pasadena, CA respectively, how do users access both massive datasets to do their analysis? EOSDIS, again, is leveraging the cloud to solve some of these issues - by storing both datasets in the same cloud region (a named set of AWS resources in the same geographical area [2].) Users wanting to access one or both of these science products will have, essentially, unlimited compute resources (or as much as can be afforded) at their fingertips.

While not all of the EOSDIS datasets are currently in the cloud, future missions, specifically larger ones, might also be located in the cloud. There is also a process for moving synergistic datasets from on-premises DAACs to the cloud, so that users can take advantage of the storage, network, and compute available in this environment.

End User Capabilities and Access

Another challenge directly impacting multidisciplinary data access are the services and interfaces available to the end user. Previously, DAACs built their own tooling and access which worked well on a DAAC-by-DAAC basis, but this led to complication and frustration from the user community when accessing multiple archives. While the cloud does not completely resolve this problem, it allows EOSDIS to focus on consistent interfaces, access mechanisms, and software reuse across DAACs to help mitigate these issues moving forward.

The EOSDIS focus on reusability and common software also means the DAACs can concentrate on building tools and services to better enable the end user science communities. This means transformations of raw data into user preferred formats, matching target datasets' projection and/or resolution, and some high level analysis (e.g. time series) that can be run on a given geographic region.

While transformations and services are important on the archived data itself, the cloud also gives the opportunity to explore more analysis-ready data (ARD) formats. Imagine having the power to define a complex set of criteria (quality flag, land mask, etc) and then perform rudimentary to complex calculations on that filtered data. Now imagine all that happening in real time - an interactive science browsing system. DAACs are also exploring services that enable cloud-optimized formats, which can be leveraged by the end-user that wishes to perform analysis-in-place (in the cloud) most efficiently. These are the next generation services the DAACs can offer, while maintaining the rich metadata, citation, and collection information users have grown accustomed to.

Potential Barriers to End User Adoption

While ESDIS and the DAACs are excited about the cloud based future, they are also very much aware there are a number of issues that may impact users and their adoption of the cloud. This section calls a few of these out by name, as well as acceptance and mitigation standards used to address the core issue.

Education and Skill Sets Required in the Cloud

Proponents of moving to the cloud often cite the simplicity of cloud solutions. The idea of focusing on your domain while letting AWS or Google do the heavy lifting sounds enticing, but is far from the truth. In order to leverage cloud functionality (scalability, elasticity, infinite storage), most applications and scripts will need to leverage new services and APIs. For instance, to get a durable (i.e. file loss resistant/tolerant) data store, one needs to move away from "file based" systems and migrate to something like Simple Storage Service (S3), which is an "object storage" system. If your processing is heavily file dependent (I/O), there may be some growing pains with moving files around and vice versa.

Cost is another aspect to consider. Many users cite economic understanding and costing of the cloud as a barrier to usage. Is there a good way of understanding and forecasting cloud budgets? And how does this change the more and more you utilize cloud services offered? For more on cost in the cloud, see the **Understanding and Managing Costs in the AWS Cloud** tutorial.

Economic Situations not well-suited for Cloud Use

There are some situations where moving to the cloud may not be ideal. Users' processes that need to store and egress (e.g. download from the cloud to a non-cloud machine) a lot of information for a long period of time may not be the right fit. The cost models of storing data usually become more expensive over time compared to purchasing your own hardware for storage, especially if you expect to use the same hardware for 5 years or more. Networking and computing resources may behave differently depending on the usage patterns of your hardware.

Hardware such as a processing server that is consistently and constantly being used (e.g. CPU or memory usage greater than 80%) may be better suited for purchase and running locally since the cost of running this instance in the cloud gets expensive and you lose the tradeoffs of scalability and elasticity (you lose these because your processing level is consistent - it's not fluctuating up and down drastically, so being able to add or remove CPU does not help).

Access to the cloud

It is quite possible that some organizations that may not be permitted to utilize selected cloud computing resources (i.e., those cloud resources that EOSDIS is utilizing - e.g., AWS). Some educational institutions may direct users to local computing facilities, while other government agencies may not have the infrastructure and accounting in place to allow access in the near- term. Finally, partner organizations overseas have shown reluctance and hesitation to use the cloud

for different reasons. No matter the reason, EOSDIS and the DAACs will continue to serve end users on all platforms, including the cloud, supercomputers, and on-premises solutions.

To the Cloud: DAAC Considerations

Storage

From the DAACs perspective, the amount of infrastructure to store this data is exceptional. Not only do we need to store the data for the life of the mission, but in perpetuity as well. Add to that redundancy (multiple copies of the data in case of disk failures) and disaster recovery (the ability to restore an entire archive should something catastrophic happen), and we're looking at storing multiple copies of multi-petabyte data! All of this would require racks and racks of servers for data storage. Servers need to be maintained and replenished on regular schedules, and doing this at the scales of Big Data often requires teams of engineers to support them. This support gets away from the core purpose of the DAACS: to archive and provide access to science data products.

By moving to the cloud, we treat storage as a commodity—if more storage is required, DAACs can quickly provision and pay for more. If less storage is needed, we remove resources and pay for less. The DAACs have greater flexibility to move data to cheaper (colder) storage options when it has become less useful (e.g. newer versions have become available), while preserving the ability to access the data should the need arise.

Network

Much like storage, network capacity is a critical component for the Earthdata enterprise, both for the distribution of data to the end users and for the ingest of the science data products. To ingest multiple terabytes of data per day, DAACs need to access that data and pull it from the provider's location (or have the provider push the data). While a single large mission might be feasible in an on-premises network, what happens when another big mission comes in? Can we easily double the amount of network capacity we have if a second mission needs massive amounts of data throughput?

In fact, there doesn't even need to be a second mission—a reprocessing campaign (when older data is re-processed through newer versions of code, algorithms, and ancillary products with greater accuracy or improvements) by the mission itself can be run at orders of magnitude higher than the forward processing rate. To ensure the capacity to handle this in an on-premises solution, DAACs would need to procure (and pay for) the maximum bandwidth volume they expect; but would only use all of this bandwidth during those specific times of reprocessing; often only once a year.

References

[1] https://www.ibm.com/cloud/blog/ibm-data-catalog-data-scientists-productivity

[2] https://docs.aws.amazon.com/general/latest/gr/glos-chap.html