Landsat in the Cloud

A Report of the National Geospatial Advisory Committee
Landsat Advisory Group
May 2022

Table of Contents

Table of Figures

1.0 Executive Summary

The U.S. Geological Survey (USGS), in early 2021, requested the Landsat Advisory Group (LAG), subcommittee of the National Geospatial Advisory Committee (NGAC) to provide guidance on what innovations its National Land Imaging program might consider when providing access to and use of data and information products in the cloud. The task team examined several case studies from public and private organizations who made the decision to use cloud services and has included aspects of those learned experiences in this paper.

From that study, several recommendations emerged:

1. Select a single commercial cloud provider to host the “master” publicly-accessible copy of Landsat data (i.e., those collection products intended for release to the public).
2. Utilize commercial cloud features such as replication of data into multiple availability zones (separate cloud data centers in separate geographical regions) to ensure a high level of resiliency.
3. Provide a mechanism for this data to be accessed by users on other cloud platforms.
4. Where possible, negotiate with the cloud providers for USGS to pay reduced or no cost for the data hosting and data egress.
5. Conduct further analysis of the appropriate economic model for providing this data, particularly with respect to the distribution of egress and cloud compute/ analysis costs among the government and users, to ensure that these choices are consistent with Landsat’s free and open data policy.

Within this white paper are the background explanations that led to these recommendations. Each recommendation has additional explanation that should be considered as part of the overall guidance.

2.0 Introduction

Over the past decade, a rapidly growing number of companies and public entities have migrated their operations to commercial clouds, motivated by factors of resiliency, scale and economics. They have also been motivated by “network effects”: as more users themselves move to the cloud, sharing and large-scale analysis of data is easier. In the cloud environment, it is possible to pass a “pointer” to the data rather than copying it from place to place, making it even more attractive for new users to adopt use of the cloud. Commercial cloud providers offer servers, storage, and virtual desktops, along with a plethora of pre-build services that reduce the amount of unique development required and speed time-to-operation, or time-to-market for commercial firms.

This LAG report examines key decisions involved in the transition to data provision in the cloud environment. The study team provides recommendations for how USGS can best make use of public cloud infrastructures, leveraging the lessons learned by both public and private entities who have themselves made the shift to the cloud. A combination of growing volumes of Landsat (and other remote sensing) data, along with applications that are increasingly performing analysis of this data at country or global scale, along with deep temporal time series, makes the scale capabilities of public clouds increasingly attractive for users, and we believe, in turn, for USGS: the cloud makes Landsat data even more accessible to and analyzable by the user community, while facilitating more efficient data management techniques and furthering the promise of open data, and should do so at a cost that benefits the taxpayer.

3.0 Characteristics of the Cloud

As most these days are likely already familiar, cloud-based infrastructures have some similarities with legacy data systems. In both cases, the infrastructure includes servers that organize, store, and process data as well as user interfaces or application program interfaces (APIs) accessible via the internet that allow potential users to find and access data, with several key differences:

1. Cloud-based systems are shared across many users, rather than each user having its own data system / data center.
2. As a consequence, cloud-based systems are designed to allow users to analyze and work with data “in the cloud” without copying (downloading) the data to their own local system.
3. Cloud-based systems, particularly commercial clouds, operate at enormous scale, which unlocks the potential to perform data analysis at much larger scale than would be affordable in a dedicated data center. This is discussed further below with respect to economic considerations.
4. Cloud-based systems also offer options for leveraging a broad range of commercial cloud services beyond storage and processing, including robust cybersecurity features. These services and associated toolsets, often only available in the cloud and not available for on-premises solutions, can significantly reduce time and effort to develop and deploy applications.

Ranked among the leading commercial cloud service providers in 2022 are Amazon Web Services (AWS), Microsoft Azure, Google Cloud, Alibaba Cloud, Oracle, Salesforce, SAP, Rackspace Cloud and VMWare. Some large companies use multiple clouds to meet unique needs for different aspects of their business whether file storage, messaging, database, developer tools, documents, or even email. As an example, Maxar (a commercial remote sensing data provider) began its cloud transition in earnest in 2014, driven by the need to operate a larger satellite constellation (a scale challenge), to produce data in larger volume for more use cases such as AI/ML analytics (another scale challenge), and to make the imagery archive more widely available on a more-timely basis to Maxar’s customers, many of whom were already using AWS. As of the present, all of Maxar’s public imagery archive (>120 petabytes) has been moved to the AWS commercial cloud.

4.0 Private Cloud vs. Commercial Cloud

As noted above, one of the opportunities offered by the transition to cloud infrastructure is the ability to leverage existing commercial cloud services.[1] These entities manage cloud infrastructures much larger than those needed to accommodate satellite data. There are pros and cons to working with these entities. Doing so requires that entities give up some level of control over the design and operation of the system. However, there are efficiencies and cost-savings associated with the use of these systems. Almost all entities[2] that we examined determined that the benefits offered by utilizing commercial cloud providers outweighed any downsides.

In 2012, NASA shut down its internal Nebula cloud computing system based on the results of a 5-month test that benchmarked Nebula's capabilities against those of Amazon and Microsoft. According to NASA, "The test found that public clouds were more reliable and cost effective and offered much greater computing capacity and better IT support services than Nebula."[3,4] NASA now works exclusively with AWS, but could move to a multi-cloud model in the future.

During the first few years of its cloud migration, Maxar focused on in-house cloud as a means to ease the transition, but experience showed that this was a mistake in that it slowed Maxar's transition and delayed the learning curve for its team members. If Maxar had the opportunity to revisit this decision, it would have chosen to go directly to the commercial cloud.

The remainder of the paper focuses on the benefits, challenges, and specific issues associated with utilization of commercial cloud infrastructure.

5.0 Benefits of Cloud Systems

Cloud systems, particularly those provided by commercial entities, enable rapid development and scaling to accommodate the needs of data providers and users. They can help to improve access to data among new types of users as well as those who lack funding to acquire or operate their own large data centers. They improve efficiency and speed of access. They also enable users to collaborate throughout and across geographic areas, and, depending upon whether they are intended for IaaS (Infrastructure as a Service), PaaS (Platform as a Service), or SaaS (Software as a Service), they facilitate sharing of data, data operations, and analysis.

[1] Note that Cloud Services models include Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS).

[2] One exception to this is ESA. Although some of ESA's satellite imagery including Sentinel have been made available on AWS, ESA started to build its own private cloud within the last decade. These implementations include computing facilities, storage and back-up solutions, IT security and protection measures in line with the corporate security policies and directives.

[3] NASA Office of the Inspector General. NASA's Progress in Adopting Cloud-Computing Technologies. REPORT NO. IG-13-021. https://oig.nasa.gov/docs/IG-13-021.pdf

[4] See: NGAC Paper Evaluation of a Range of Landsat Data Cost Sharing Models June 2019 at https://www.fgdc.gov/ngac/key-documents and NASA Office of Inspector General Report No. IG-13-021 (Assignment No. A-12-022-00) https://oig.nasa.gov/docs/IG-13-021.pdf

5.1 Rapid Development

Maxar found that improved development velocity occurs, both because using existing cloud services avoids the need to develop them internally, and because there is no lead time required to provision hardware capacity to roll out a new product or service (unlike in the data center world where one must acquire, install, and tune new hardware and any specific data management and operating systems).

5.2 Resiliency

Cloud providers offer a range of capabilities to ensure both high uptime and to protect against loss of data. Examples include the ability to provision services from and/or replicate data in multiple geographic regions, as well as highly reliable storage tiers for long term archives. When Maxar migrated to the AWS cloud, it was able to retire its legacy on-premises tape libraries, achieve superior resiliency, and do so at a net cost savings relative to its legacy approach.

5.3 Rapid Scaling

Maxar also found that the cloud allowed for rapid auto-scaling of capacity: increasing capacity to accommodate growing data volumes or decreasing capacity to allow for cost management. They also noted that the cloud environment provided the ability to work on very large-scale projects (e.g., country- or continent-scale) with constrained timelines because the entire archive is online and available in the cloud, and parallel computation can be employed to apply AI/ML algorithms at massive scale rather than being limited by the capacity of an on-premises data center.

5.4 Cost

The ability to scale also resulted in cost savings for Maxar, since it did not have to purchase/refresh on-premises data center capacity that envelopes the upper bound of usage and is acquired in advance of date of need. NASA also noted the financial efficiency of the cloud, as NASA only pays for storage and services actually used.[5] This allows the amount of storage or services to be continually adjusted to ensure that data and services are effectively provided at the lowest possible cost to NASA. The cloud system also has the potential to decrease costs of data provision and/or access for users. The considerations involved in this element are discussed in detail below.

[5] Cassidy, Emily. "Cloud Data Egress: How EOSDIS Supports User Needs Share on Twitter Share on Facebook Share on Pinterest." NASA. https://earthdata.nasa.gov/learn/articles/cloud-data-egress. Accessed March 28, 2022.

5.5 Flexibility

NASA notes that mission needs can dictate options for selecting operating systems, programming languages, databases, and other criteria to enable the best use of mission data.[6]

5.6 Improved Access

As mentioned above, the cloud enables users to work with the data at any scale, regardless of the capabilities of their local systems. Users that don't have access to significant storage capacity or big data management or advanced computing technology in-house can now access these capabilities within the cloud. This can help to reduce inequalities among researchers with different institutional capabilities, allowing a further expansion of the use of satellite data by a wider range of users. (As noted below, however, researchers still require funding to pay fees to carry out analysis in the cloud.)

To deliver next-generation geospatial services to users, Planet moved all of its satellite imagery and image processing to Google Cloud Platform as part of a multi-cloud strategy[7].

NOAA has found that with its data located in the public clouds, users have ready data to integrate with other data, collaborate projects, and share data, codes, and findings in the clouds. The cloud solution not only enhances data access but drives applications and economic values of NOAA's data. NOAA's three commercial cloud partners offer extensive data access and reduce the users' need to download the data and can analyze the data next to the computational capability, or in some cases, can pre-load into analytic tools in the cloud. Enabling NOAA data in the cloud significantly reduces the effort needed to work with such large data volumes. This in turn encourages the user to continue exploiting rich NOAA observations and reanalysis data and crowd-sources talents to accelerate knowledge production from NOAA data assets.

5.7 Speed

User access to data is also rapid - there is no need to undergo the often time intensive process of downloading the data. NASA noted that this can further expand use of its extensive data collections by improving ease of access. Improving data access in the cloud can also increase the speed at which data users can carry out analysis, while intensifying the search for key elements of information. NASA noted that "data users can bring their algorithms and processing software to the cloud and work directly with the data in the cloud, simplifying traditional procurement procedures for hardware support while expediting science discovery."[8]

[6] NASA Earth Data. "Earthdata Cloud Evolution." NASA. https://earthdata.nasa.gov/eosdis/cloud-evolution. Accessed March 28, 2022.

[7] https://cloud.google.com/customers/planet

[8] NASA Earth Data. "Earthdata Cloud Evolution." NASA. https://earthdata.nasa.gov/eosdis/cloud-evolution. Accessed March 28, 2022.

The cloud also enhances collaboration, making it easier for geographically distributed teams to coordinate.

5.8 Increased Data Usage and Reduced Demand on Government

NOAA reported that its old-fashioned tape-based archives quickly reached their limit with massive data volumes and extensive data demands. In 2015, NOAA launched its Big Data Project[9] as a pilot program to examine the possibility of using commercial cloud providers to distribute its data. It later operationalized this effort as the Big Data Program, engaging in partnerships with Microsoft, Amazon, and Google in early FY19. Ansari et al.[10] summarized the early success of NOAA Big Data efforts. From 2015 and 2016, the cloud-service alliance received 270 TB and 180 million files of NOAA data. AWS publicly released access to Level II NEXRAD data on 27 October 2015. From November 2015 to July 2016, users' data-access volume totaled 445 TB. Requests for NOAA NEXRAD data reached 94 TB in March 2016 (compared to 18TB in March 2015). NOAA's National Center for Environmental Information (NCEI), responsible for non-cloud NEXRAD data dissemination, reported an 84% reduction of data requests from January 2015 to July 2016. NOAA's Big Data Program enjoyed greater and faster data access by broader users and higher productivity within the institution and across extensive industries.

In the case of NOAA’s Big Data Program, in addition to cloud computing and free and open data access, the industry provided data services to commercial entities for application developments with NOAA data in transportation and agriculture, as two significant examples. An increase in public access to both data and computing facilities in the cloud enhances opportunities for innovative applications of NOAA data which may otherwise remain untouched in government archives.

5.9 Enhanced Public-Private Partnership, Collaborative Innovations, and Data Economic Value

NOAA’s Big Data Program leverages the industry's expertise in data storage and access and NOAA's data archives and subject matter expertise. With NOAA’s data in the cloud, users have ready data to integrate with other data, collaborate on projects, and share data, codes, and findings in the cloud. The cloud solution not only enhanced data access but drove applications and economic values of NOAA’s archived data through effective public-private partnerships. The ability to collaborate and share in this way can also reduce redundant tools and services and enforce the use of community standards as well as uniform policies and processes.

[9] https://www.noaa.gov/information-technology/big-data

[10] Ansari, S., Del Greco, S., Kearns, E., Brown, O., Wilkins, S., Ramamurthy, M., ... & Lakshmanan, V. (2018). Unlocking the potential of NEXRAD data through NOAA's Big Data Partnership. Bulletin of the American Meteorological Society, 99(1), 189-204.

6.0 Challenges of Cloud Systems

While cloud systems provide many benefits, they also pose challenges[11]. The transition to the cloud requires deliberate effort on the part of the entity undergoing the transition, and will be most successful when both the technical and economic design of the system are carefully considered, as described in the sections below. In addition, while the number of users familiar with this technology is rapidly increasing, many existing data users may need to undergo training to adopt the new system.

As an example, Maxar's journey to the cloud was not without challenges. There was a learning curve for its technical staff, and given the rate of change in the cloud environment, it required continuous oversight. Given its massive data management challenges, Maxar intentionally has been on the leading edge of cloud usage. As a frequent beta customer, it worked with the cloud provider (AWS) to identify defects or areas for improvement, and often worked with incomplete versions that lacked features that would be in the "next release," as well as compute instance type[12] availability. Lagging behind the early adopter curve by 2-3 quarters would have reduced these impacts but at the expense of delaying the cloud implementation; on balance, the small pain of being a beta customer was outweighed by the larger benefit of speed of implementation.

7.0 Additional Considerations

7.1 Changes in Cost Structure

Migrating from a legacy on-premises data center to a public cloud requires several paradigm shifts. First and most significant among these is economic; in the on-premises data center model, the costs to acquire, maintain and operate the data center are largely fixed, whereas in the public cloud, these are variable and scale with usage of storage, compute, and bandwidth utilization. As noted above, this means that the system can (and should!) be designed to minimize costs within the cloud infrastructure, and this often requires re-architecting of systems that were previously designed to operate in an on-premises data center. That means company-or organization-wide budgeting for and managing cloud resources as an ongoing process that recognizes the newly-introduced business model requiring the appropriate technical expertise, the tools, and focused oversight and governance.

Because the cloud is a pay-for-use environment, the costs of operations are more visible, but only if they are constantly monitored to avoid unnecessary costs with over-provisioning storage or neglecting to turn off cloud instances when they are no longer needed. What was previously "below the line" (buried in the IT budget) is now visible in monthly invoices, and this drives a

[11] In May, 2021, the GeeksforGeeks identified what it currently found to be the most common cloud computing challenges: data security and privacy; cost management; multi-cloud environment; performance challenges; interoperability and flexibility; high dependence on network; and lack of knowledge and expertise. https://www.geeksforgeeks.org/7-most-common-cloud-computing-challenges/

[12] Instance types comprise varying combinations of CPU, memory, storage, and networking capacity and give you the flexibility to choose the appropriate mix of resources for your applications.

need for a cloud governance competency. Maxar concluded that on balance, this was a net positive in that it gives Maxar much better insight into the true cost of a product or service, and has enabled a culture of continuous improvement that in turn drove architectural or operational changes based on usage data that was previously invisible (or less relevant) in the legacy on-premises data center.

This highlights another important aspect of cloud migrations. While on an apples-to-apples comparison, migrating a service to the cloud can often result in a cost reduction, utilizing the greater features and capabilities offered by the cloud can result in an increase in absolute cost. The highly granular visibility into costs offered by cloud providers makes it possible for the organization to evaluate when the value offered by the capability justified its cost. This visibility gives the enterprise, whether private or public, the necessary insight to perform cost/benefit trades.

7.2 Increased Efficiency via Cloud-Native Design

Transitioning to the cloud requires both changes to software to more efficiently (and cost-effectively) utilize the cloud. A lesson learned shared by multiple entities was the importance of designing a “cloud native” infrastructure, rather than undergoing a simple “lift and shift” of legacy applications, i.e., porting them to the cloud without any changes to the applications). This type of redesign typically involves having applications that only run when they are needed, having the most frequently used data stored in more quickly-accessible formats than others, and optimizing the type of computer hardware instances. Commercial cloud providers have also introduced more hybrid on-ramps to make it easier to migrate to the cloud.

There are several reasons re-optimizing applications for the cloud environment can result in a more efficient (and cost-effective) solution than simple “lift and shift.” This paper does not go into complex technical detail about best-practices for this process, however, there are numerous case studies online from a range of cloud providers, application providers and users that provide useful insight into planning the most efficient cloud migration strategy[13]. This paper makes the following observations associated with evolving best-practices.

Pay-by-Use: In an on-premise data center, an application can typically be left running regardless of whether it is being used or not, but in the cloud the “meter is always running.” Re-architecting system use, so that the application is only running when used, avoids this running expense, and moreover allows the application to scale to spikes in demand without the cost of

[13] [List of web references to case studies on cloud migration]
https://aws.amazon.com/blogs/enterprise-strategy/cloud-native-or-lift-and-shift/
https://cloud.netapp.com/blog/cloud-migration-strategy-challenges-and-steps
https://www.cisco.com/c/en/us/solutions/cloud/what-is-a-cloud-migration-strategy.html
https://increment.com/cloud/case-studies-in-cloud-migration/
https://www.cognizant.com/us/en/case-studies/aws-cloud-migration-healthcare

sizing a data center to handle peak demand or to achieve resiliency through duplication of hardware.

Data Storage and Compute Options: Cloud environments typically have many choices for compute instances and storage types, each optimized for a different purpose and each with a different price. Cloud-native applications take advantage of this to minimize cost, e.g., by allocating non-time-sensitive background compute tasks to currently unused compute instances that are lower cost but are interruptible when a higher-paying use comes along (think of this as “space available” pricing), or by dynamically migrating infrequently used data to lower cost but higher latency storage instances. Cloud providers typically provide a number of services to help automate this type of optimization.

Data Migration: Moving data, and often the associated workloads, to a cloud must strategically address management and technology challenges because company resources - including the affected staff and workforces - and assets - like existing data centers – may dramatically impact the complex environment of the business (whether for-profit, not-for-profit, government, or academic) itself. Each organization must decide whether all or only some data is migrating from the on premise system to the cloud. Sometimes the platform of choice may involve hybrid or multi-cloud architecture. Migration planners face several options to move data from an on-premises data center to the cloud.

Maxar performed its data migration with a mix of AWS Snowmobile for most of its historical data (This enterprise was the first customer of this “data center on wheels”) along with a dedicated AWS direct connect (network link) to its internal network to deliver new data to the cloud. Snowmobile was by far the fastest way to move what was at the time roughly 100 petabytes of data to the cloud, and far more cost effective than any form of network connectivity.

Not all migrations were initially successful. As an example, National Geographic began by using the Nirvanix Cloud for its archive and distribution[14]. They needed to develop a specific situation where the upload to the cloud could not be computer to cloud but from old media like tapes and discs to cloud. Nearly a decade ago, Dan Backer, then Director of its Infrastructure Systems, said National Geographic would not close its Washington-based data center, but would move to the cloud to allow a reduction in equipment, maintenance, and power use. Although Nirvanix financially failed, the transition to the cloud storage, maintenance, and use technology model prevailed for the organization, confronted by another migrations step and now served by AWS. Within this past year, the organization, building upon the success of expanding its use of cloud services, announced the development of a cloud-based Conservation Intelligence Platform, aimed at providing real-time identification of endangered species and security alerts for protected areas.[15]

[14] https://www.computerworld.com/article/2502458/national-geographic-moves-media-archive-to-cloud.html
[15] https://www.nationalgeographic.com/impact/article/corey-jaskolski-conservation-technology

Moving to the cloud requires commitment to have a solid plan with expert leadership. A company outside our remote sensing and geospatial domain but with a massive data store, over 14,000 branch offices, and more than 7 million clients, Edward Jones, in 2022, has been recruiting for a Test Automation Architect, a Cloud Architect, to lead its "Quality cloud" migration strategy.

Utilize Existing Cloud Infrastructure and Tools: It's not uncommon when migrating to the cloud to be able to take advantage of services and applications that are considerably less costly than legacy licensed application software (e.g., databases), including a growing set of open source software that has been fully integrated into the cloud environment with a set of tools to facilitate migration. Many achieve significant savings on annual software license costs with a move to the cloud.

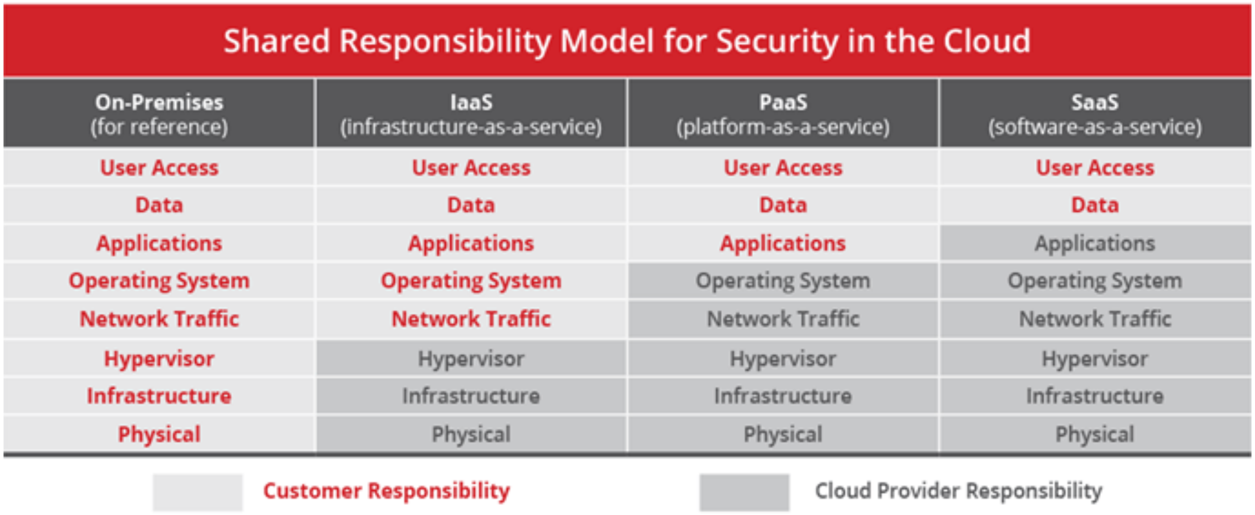
Maxar reports that it uses cloud native tools provided both by Amazon and third parties. Of relevance to this paper's recommendations later, incorporating information from the cited case studies: Maxar pays for its cloud costs but depending on product/service offering will pass those costs on to its customers; for example, pricing for Maxar's online SecureWatch product is based on a $/gigabyte model.

7.3 Security Considerations

A cloud is a shared infrastructure. While cloud providers provide a robust set of tools for controlling which user has permission to access which data or other resources, as well as a very broad set of cybersecurity tools, it is essential that these tools be utilized. The cloud is arguably more capable of achieving high levels of security than individual data centers, but in effect it is a software-defined data center so errors in configuration can expose data or intellectual property more quickly, as well as incurring unintended costs, thus cloud users must make use of the tools available to them. Taking a cybersecurity focus from the beginning on the plan to move to the cloud environment is crucial. While the Federal government can only utilize FedRamp-compliant tools, most cloud vendors continue to gain approvals for the most commonly used tools.

McAfee[16] provides a good infographic that illustrates shared responsibilities. "The cloud service provider covers security of the cloud itself, and the customer covers security of what they put in it. In every cloud service—from software-as-a-service (SaaS) like Microsoft Office 365 to infrastructure-as-a-service (IaaS) like Amazon Web Services (AWS)—the cloud computing customer is always responsible for protecting their data from security threats and controlling access to it.

[16] https://www.mcafee.com/enterprise/en-us/security-awareness/cloud/security-issues-in-cloud-computing.html



Shared Responsibility Model for Security in the Cloud

On-Premises (for reference)	IaaS (infrastructure-as-a-service)	PaaS (platform-as-a-service)	SaaS (software-as-a-service)
User Access	User Access	User Access	User Access
Data	Data	Data	Data
Applications	Applications	Applications	Applications
Operating System	Operating System	Operating System	Operating System
Network Traffic	Network Traffic	Network Traffic	Network Traffic
Hypervisor	Hypervisor	Hypervisor	Hypervisor
Infrastructure	Infrastructure	Infrastructure	Infrastructure
Physical	Physical	Physical	Physical

□ Customer Responsibility ▨ Cloud Provider Responsibility

Figure 1: *Shared Responsibility Model for Security in the Cloud*

7.4 Data Integrity

One concern with regard to cloud systems is the need to ensure that data integrity is maintained. Data producers want to ensure that users have access to the most accurate version of the data as well as the relevant metadata and other information necessary to use the data correctly.

In its Big Data Program, NOAA and its cloud services providers abide by the three general rules below:

1. Data must be available for public use without restriction.
2. Data must still be archived following applicable federal data directives.
3. Data must have a data owner/SME to support transferring the data and user questions regarding the data.

Under the public-private partnership, NOAA BDP assures data quality and provides data expertise to support cloud service providers' data access. Cloud services providers consider NOAA data expertise the most valuable resource in the partnership and afford those cloud providers on-going opportunities to stay involved with NOAA and further collaborations. Cloud services providers also utilize NOAA data to build tools and capabilities for business sustainability metrics for their business operations and customers. The comparable potential of Landsat data to improve local to global sustainability measures attracts the same cloud services providers as already evidenced in the USGS efforts to migrate to a cloud environment.

NOAA's Cooperative Institute for Satellite Earth System Studies at the North Carolina State University serves as a data broker who overcomes barriers and connects disparate cultures between NOAA and cloud services providers. The data broker facilitates data transfers and cloud mechanisms and addresses experimental infrastructural needs. For example, the data broker coordinates publishing NOAA data from the federal systems to all three cloud platforms and serves as the single point of contact for the cloud services providers to simplify operations and security concerns. The USGS EROS Center would act as a Landsat data broker in a similar scenario.

7.5 Free and Open Data in the Cloud

When data is made available in the cloud, the new infrastructure, distribution of costs, and variable cost structure raises questions about the meaning of “free and open data” within this environment. In this paper, the questions are pertinent to federal government collections of remotely-sensed data and government produced or contracted- for derivative products. In the past, with taxpayer funding and Congressionally-approved budgets, federal agencies have paid for the acquisition, processing, storage, and provision of data and data products. As web access and agencies’ computerized data management systems were evolving, most users were permitted to access or download the data without paying.[17] (Subsequently legislation, rather than just policy[18], has issued requirements about “open” data with respect to accessibility.)

The legacy systems’ costs were largely fixed within the agency’s operating budget. It was not readily apparent what cost was associated with an individual user’s activities. Also, the performance of the system was similarly fixed, e.g., the rate at which data could be downloaded was limited to the bandwidth provided by the legacy system, which fixed costs but limited the ability to scale.

Similarly, users paid the costs associated with analyzing downloaded data, but this was a fixed cost associated with their personal computing systems or those owned by their home institution. Unless some internal accounting system was aligned with tracking various computer transactions, users did not face a per-project cost for data manipulation and information analysis.

These things have changed in the cloud environment. If a user moves a copy of the data outside of the cloud, commercial cloud providers monitor and charge for the data egress (download) activities for that specific user. If analysis is conducted within the cloud environment, the compute activity of each user is similarly monitored and charged for.

This pay-as-you go configuration poses a challenge for government agencies committed to freely available and openly accessible data. Most agencies accept that the government agencies must still pay the cost of data storage within the cloud environment. However, does free and open access to taxpayer-funded data and products require that governments also cover the costs of data egress and/or analysis? If they do, this requires a new approach to budgeting, with

[17] Until 2008 Landsat was an exception to the practice that imagery obtained by federally funded Earth Observation satellites was provided to users typically at no cost. The United States Geological Survey (USGS) made Landsat data accessible via the internet for free at that time.

[18] https://www.congress.gov/115/bills/s760/BILLS-115s760is.pdf

the development of estimates for these costs. Further, if these costs are born by the agency, a technical limit must be in place to ensure that costs do not go above the budgeted amount.

If agencies do not cover these costs, users are not able to use the data without paying an upfront cost (either for egress or cloud compute capability). The spirit of the "free and open" data policy has historically been to ensure that people, including – or even especially – those with few resources, can access and use Landsat data without paying any fees. Previous research has shown that this ability to access data at no cost has been essential to generating the significant socioeconomic benefits associated with these policies.[19]

If agencies choose to cover one of these costs but not the other (e.g., egress costs, but not cloud compute capability), they risk creating perverse incentives for users. For example, if egress if free to users, but cloud compute is not, then users will be incentivized to remove data from the cloud, even if it would be cheaper and more efficient to conduct their analysis in the cloud environment. Such inefficiencies could ultimately cost the government more than if both options were subsidized (up to some yet-undefined point that will require more analysis and understanding a variety of use cases).

An additional challenge is that the shift to cloud computing is a structural change affecting not only government agencies, but actors throughout society. For example, in the past, many individuals were issued a laptop as part of their job. Even those without a computer could typically access one for free at a library or other community location. This is not the case with respect to cloud compute capability. Even universities, which may be expected to be early adopters of cloud technology, do not typically issue free cloud compute capabilities to their faculty or students. These types of gaps in the new economic structure associated with cloud computing must be carefully addressed, both during this transition period and in the long term.

At present, government agencies have taken a variety of approaches to addressing these issues. NASA chose to cover costs (up to a specified point) associated with all of these elements. With regard to NASA data, "a user will be able to employ these cloud-based EOSDIS-provided services to discover, search, access, and download data, at no cost to the user. The user has the option to:

1. Download the data to their own local machine, server, or HPC account – free of charge;
2. Connect to the Earthdata data from their (AWS) cloud instance – free of charge; or
3. Move the data out of the AWS-hosted cloud (free of charge) to another cloud platform of choice."[20]

Since the cloud is pay-as-you-go the EOSDIS team had to build new tools to ensure that the system can not violate the Antideficiency Act. This law prohibits the government from spending

[19] Borowitz. "Open Data: The Global Effort for Open Access to Environmental Satellite Data." 2017.
[20] https://earthdata.nasa.gov/learn/articles/tools-and-technology-articles/ghrc-moves-to-the-earthdata-cloud

money that has not been appropriated.[21] If a user wishes to store data in their own (AWS) cloud instance or cloud storage, the user is responsible for that storage cost."[22]

By contrast, ESA has adopted a system in which data view, download, and initial processing (visualization, data cubes, clip, etc.) is free to users. However, anything that utilizes storage, GPU/CPU or virtual computing requires payment per usage.

Working under the 2018 Federal Cloud Computing Strategy, in March 2020, the USGS placed a copy of its consolidated Landsat global data inventory into a commercial cloud. At that time, Landsat's cloud architecture began using a hybrid approach with both the USGS cloud hosting solutions program and existing EROS center private cloud capabilities. USGS officials then announced that all current users of Landsat data, including redistributors, would continue to have access to Landsat data at no cost for downloading and processing files[23]. No cost processing, however, may not be unlimited, even using open-source tools. A recent video prepared by USGS EROS Center notes that some processing requests may incur costs. [24]

The NOAA Big Data Project was implemented via Cooperative Research and Development Agreements (CRADAs) with Amazon Web Services, Google, and Microsoft Azure at no cost to the government and was operationalized in Q1 FY19[25]. The data alliances provided full NOAA data access open and free to the public and industry but charged use fees for virtual machines on their IaaS. Without these public-private partnerships, egress costs for NOAA data would increase along with user demand, and NOAA felt that funding those rising costs into the future would be prohibitive. The BDP partnership ensures all cloud users can egress NOAA data hosted by that public cloud provider for free. In addition, some clear distinction about the data on the cloud must be understood. "The BDP Cloud Service Providers (CSPs) are tasked with distributing the original data content only. Any use of value-added products on the CSP's platform that use NOAA data, can incur a charge for those products in the same way that they do today."[26]

It is not clear what effect these different policies have had on data access and use among the various agencies, nor what effect they may have in the long term. This is an issue that requires additional consideration and study to ensure that the United States continues to benefit from the socioeconomic value associated with Landsat's free and open data policy. Recently the EROS User Group received a request from the EROS User Services, expressing interest to hear from its diverse list of members about any "newer capabilities such as you own cloud-based analysis workflows to other approaches enabled by Cloud Optimized GeoTIFF (COG) format and

[21] https://earthdata.nasa.gov/learn/articles/tools-and-technology-articles/ghrc-moves-to-the-earthdata-cloud
[22] https://cdn.earthdata.nasa.gov/conduit/upload/14964/02_Understanding_and_Managing_Costs_in_the_AWS_Cloud.pdf
[23] https://www.usgs.gov/news/technical-announcement/landsat-data-moving-public-cloud-early-2020
[24] https://www.usgs.gov/media/videos/landsat-cloud-cogs-and-notebooks
[25] Simonson, A. Brown, O Dissen, J., Kearns, E. Szura, K. and Brannock, J. (2022) NOAA Open Data Dissemination (formerly NOAA Big Data Project/Program), Earth and Space Science Open Archive. https://www.essoar.org/doi/10.1002/essoar.10508327.1
[26] https://www.noaa.gov/big-data-project-frequently-asked-questions

Spatio Temporal Asset Catalog (STAC) records."[27] Listening to users of the varied Cloud services offers excellent opportunity to continuously improve the services in the most efficient and effective ways.

8.0 LAG Recommendations

1. Select a single commercial cloud provider to host the "master" publicly-accessible copy of Landsat data, i.e., those Landsat collection products that are intended for release to the public.[28] Among the data security and protection procedures, provide users with read-only access to this data so that its provenance is entirely under control of USGS.

2. Utilize commercial cloud features such as replication of data into multiple availability zones (separate cloud data centers in separate geographical regions) to ensure a high level of resiliency and data availability.

3. Provide a mechanism for this data to be accessed by users on other cloud platforms. Initially, this may simply be enabling data to be accessed from another cloud platform, but based on usage patterns, this may evolve to providing a copy of some or all Landsat data (the same data as mentioned in point 1 above) on one or more additional cloud platforms as was done by NOAA. The economic model for this would follow the recommendations outlined in #5 below.

4. Where possible, negotiate with the cloud providers for USGS to pay reduced or no cost for the data hosting and data egress, making the argument that hosting the data in a given commercial cloud will drive demand by users for that cloud provider's services; this model was used with some success by NOAA's Big Data Program during its initial phase, though it cedes some control to the commercial cloud provider over what is and is not hosted, so unless the cloud provider agreed to host the entire gold copy, there would need to be a hybrid of on-premises (USGS/EROS) and commercial cloud. This may be a cost-effective on-ramp to full utilization of one or more commercial cloud providers.

5. Conduct further analysis of the appropriate economic model for providing this data, particularly with respect to the distribution of egress and cloud compute/ analysis costs among the government and users, to ensure that these choices are consistent with Landsat's free and open data policy.

[27] EROS User Services at custserv@usgs.gov, 14 February 2022.

[28] USGS/EROS may choose to retain the raw collected data in its own data center or archives as a backup, or move that also to the cloud. If it proceeds down the latter route, it will be important to utilize the cloud's resiliency and backup features to ensure that this raw data is never lost.

Acknowledgments

This paper was approved by the NGAC Landsat Advisory group (LAG) on May 10, 2022 and adopted by the NGAC as a whole on May 18, 2022. The LAG team developing this paper included Walter Scott, Maxar Technologies (Team Lead); Mariel Borowitz, Georgia Institute of Technology (Team Co-Lead); Steve Brumby, Impact Observatory, Inc.; Roberta Lenczowski, Roberta E. Lenczowski Consulting; Vasit Sagan, St. Louis University Geospatial Institute; and May Yuan, University of Texas-Dallas.