# National Geospatial Advisory Committee – Landsat Advisory Group

## CLOUD COMPUTING: Potential New Approaches to Data Management and Distribution[i]

**Background:** Landsat imagery provides the U.S. and the world with a continuous, consistent multispectral image archive used to monitor global resources. Landsat imagery supports a multitude of applications across scientific, societal and economic domains and is an essential national asset contributing to economic, environmental, and national security interests. The amount of data being downloaded by users has grown substantially since free, open access to the Landsat image archive was provided by the USGS EROS Center in late 2008. The EROS Center has been innovative in satisfying the needs of this rapidly increasing user base. Continued innovations in data handling and distribution will be necessary to keep pace with the increasing user demand, coupled with the fact that the digital content of a data set has grown significantly since the launch of Landsat 8 (e.g., more spectral bands having higher radiometric fidelity (12 bit vs. 8 bit quantization)).

The Federal Geographic Data Committee (FGDC) requested that the Landsat Advisory Group provide advice on "potential new approaches to data management and distribution."[ii] Included in the request was specific mention of "use of the cloud" and "investigation of the many technological advances relating to 'Big Data.'" This report focuses on recommendations to USGS relative to use of cloud computing to improve access to and use of Landsat imagery.

**Introduction: Cloud Computing, Big Data and the Opportunity for Landsat** – Cloud computing services could be implemented to provide data users with an effective way to view, select, and download desired image data. Cloud computing has become a feasible option because the necessary infrastructure is mostly in place, along with high capacity network services having enough bandwidth to provide a distributed data center service with acceptable performance characteristics. Using cloud computing as an extension of a data center is a viable method to store data in close proximity to large computing resources that are available and scalable on demand. In a traditional data center, computer resources are often over provisioned to satisfy peak demand requirements. Cloud infrastructure would enable the sharing of the same resources over multiple applications, thereby providing the elasticity to allocate more infrastructure to quickly complete a computation-intensive task where the processing load varies with time. This sharing of resources over multiple tasks has the potential to reduce infrastructure/overhead/data hosting costs.

There are many cloud services, including:

- ✓ Data Storage: Short and long term data storage from a few kilobytes to 10's of petabytes. Cloud storage may be online (on spinning disk, for fast access), nearline, or offline/archival.
- ✓ Computing Resources: On-demand processing capability for small to very large jobs requiring a few CPU cores to thousands of cores.
- ✓ Application Services: Web services, map services.
- ✓ Computing center services including DNS (domain name servers), mail servers, webservers, etc.

Each of these cloud technologies has relevance to the management of Landsat data and provision of related data services. Cloud storage is relevant because the Landsat data archive is quite large, comprising millions of scenes and multiple petabytes of data, and is growing by more than a terabyte per day. The bulk of the Landsat archive is currently offline at EROS, and therefore not readily available for provision of large-scale data services; only the most recent, and/or highest demand, scenes are readily available on spinning disk. Cloud computing resources are relevant because the processing of this 'Big Data' can be computationally intensive, from orthorectification and other pre-processing steps, to more advanced derivative product generation. The historic lack of access to scalable computation over large portions of the Landsat archive has stymied the user community in unlocking the true potential of this treasure trove of data. With the advent of the cloud, it is now possible for the first time to mine the entire Landsat archive and extract an unprecedented amount of meaningful information about our planet and how it is changing over time.

As an example of what is now possible, a team of academic, government and industry researchers published a study in the Nov. 15, 2013 issue of *Science, "*High-Resolution Global Maps of 21st-Century Forest Cover Change"[1]. Based upon analyses of massive quantities of Landsat data co-located in cloud storage (e.g., 650,000 Landsat 7 scenes or 20+ trillion pixels), this study produced the first global 30m maps of forest cover and change from 2000-2012, quantifying forest dynamics related to fires, tornadoes, disease and logging, worldwide. The analyses required over one million hours of CPU and would have taken more than fifteen years to complete on a single computer. But because the analyses were run on 10,000 CPUs in parallel in the cloud, they were completed in less than four days. The study results were made freely-available to the public online, via a cloud-hosted interactive map, presenting more than a terabyte of data at full 30m resolution[2].

**Cloud Computing and Landsat: Additional considerations** – As demonstrated above, cloud computing can facilitate new approaches to the storage, viewing, analysis and public distribution of Landsat data and derived products. However, cloud computing is not particularly useful for applications where sustained computing services are required 24/7. In such cases purchasing and running an operational data center dedicated to computing, storage, and communication services will typically be cheaper. But, cloud services can be of help to an operational center in meeting data durability and system availability requirements, as storage costs are minimal and one only pays for the storage and communication services as needed.

Security is also an important consideration. There is often tension between those who advocate using the cloud approach to cut costs and increase user access versus those who are concerned that use of the cloud will expose the services, or the data served, to vulnerabilities from outside parties. (See **Recommendation 6.)**

**Some Concepts for Cloud-enabled Landsat Data Services:** There are four main ways or 'models' for making Landsat image data accessible:

1 – **Data Download** – Data is downloaded from one storage location to another and then processed at the new location (e.g., FTP download or Zip, Clip & Ship services). Currently, the USGS EROS Center provides image accessibility via this 'model' only, which does not follow the 'cloud' paradigm since the imagery is moved to the location of processing. Although there are advantages in this method for traditional imagery workflows, it results in massive redundancy in data storage and data transfer.

2 – **Interactive Online Visualization** – Imagery is pre-processed and mosaicked at various zoom levels, and cached into tiles that serve as background imagery. This is the most scalable way to serve imagery, but it can only be used as a background image and cannot be processed for analysis/information extraction. This approach is very applicable for a number of commercial organizations that want to serve base maps (e.g., Google Maps, Bing Maps and ArcGIS online), but there is no requirement for USGS to provide this service.

3 – **Interactive Online Analysis** (on-demand interactive analysis for simple requests) – Dynamic image services allow the client application to make a request and the server extracts and processes the data before returning the required imagery or product. Such dynamic image services need to run synchronously and provide an image within a few seconds, typically as the user is panning or zooming an interactive map.

4 – **Batch-processing Analysis** (longer-running analyses for complex requests, *a.k.a.* geoprocessing) – Geoprocessing services enable the user to request an application to be performed by a server asynchronously; requested tasks can be very complex and incorporate the integration of many different data sources, as defined by the client. The output may be an image, set of values, or geometric features.

**Major Recommendations**: It is recommended that the USGS EROS Center seek to leverage cloud computing in several areas, as follows:

1. **Facilitate Landsat cloud implementations by third-party cloud providers.** In order to foster innovation, the EROS Center should create a policy and framework for supporting third-party cloud providers, most importantly by providing a bulk Landsat data download capability that is timely, comprehensive, reliable, and high-bandwidth. There is precedent for this: EROS today supports bulk download via FTP and HTTP.
2. **Facilitate the implementation (by EROS and/or third-parties) of methods that provide fast and simple accessibility to imagery, such as Interactive Online Analysis (Model 3 above).** Multiple services can be defined from the same data source that will return specific products processed directly from the Landsat L1T products, such as different band combinations, imagery in 'radiance' or 'reflectance' values, or a wide range of vegetative indices. Such services will significantly improve access to imagery by enabling users of various levels of sophistication to request and receive specific higher-level data products. This concept leverages the efficiency of the cloud by co-locating data and processing; processing only takes place on demand, when the client applications make the requests to the servers, which then process the data and return only the required information.
3. **Facilitate the implementation (by EROS and/or third-parties) of methods that provide Batch-processing Analysis (Model 4 above), as there are many scientific tasks that cannot be handled by Interactive Online Analysis alone**. Tasks such as the recent global forest cover change analyses are too large in extent to be executed in a short time frame and/or can require access to massive volumes of data. There are many different such services that could be envisaged to run on Landsat data, and most of these services would require access to a large number of scene equivalents. Currently, users wanting to perform such analyses would need to first download all the required data. An optimal model would be one in which users can define the required processing to be performed on the imagery and then transmit the model to the cloud where processing can be spread across multiple CPUs. Given the need to periodically recalibrate large groups of scenes (e.g., all Landsat 8 scenes will need to be re-processed at least once in 2014), such a cloud-based batch-processing service would support efficient, timely reprocessing.
4. **The EROS Center should investigate modification of their existing Data Download (Model 1) to enable subsets of L1T products to be downloaded**. Certain types of analyses need only operate over a time-series stack covering a small geographic area of pixels. Note that there is good synergy here with implementation of **Interactive Online Analysis** (Recommendation 2), as one way to facilitate access to such image subsets.
5. **Special attention should be given to the use of open software standards when designing any future system(s) to avoid tying any of these services to proprietary software.**
6. **Although security is an important consideration, security solutions need to be streamlined so as not to slow things down appreciably and/or make things more complicated to implement.** Given that Landsat is a public dataset of broad relevance to society, it would be unfortunate if potential innovations and beneficial applications were thwarted by excessive focus on security.

## SUMMARY

The USGS EROS Center has done a commendable job in keeping up with the rapidly increasing demand for Landsat images using traditional, scene-based data management and distribution practices. However, there appears to be a need for a paradigm shift in their data management and distribution practices in order to keep pace with projected increases in user community demands and related data volumes. A top-level investigation of cloud computing was conducted, limited by time and resources, but performed by a group of individuals with up-to-date knowledge and experience relevant to this topic. More in-depth discussions of cloud computing and 'Big Data' can be found in other recently completed studies referenced below. The consensus opinion is that adaptation of the recommended cloud computing approaches is warranted at the EROS Center, whose people are charged with the responsibility of serving a very large, diverse, and growing user community.

References:

1. Hansen, M. C., et al. "High-Resolution Global Maps of 21st-Century Forest Cover Change." *Science* 342.6160 (2013): 850-853. http://www.sciencemag.org/content/342/6160/850

2. Hansen, M.C., et al., Global Forest Change Data Portal. *Science* 342.6160 (2013): 850-853. http://earthenginepartners.appspot.com/science-2013-global-forest

3. The Magellan Report on Cloud Computing for Science, US DOE Office of Advanced Scientific Computing Research (ASCR) December 2011

4. NOAA CLASS Cloud Computing White Paper, March 2011, Document ID 1340.

5. Redefining the possibility of digital Earth and geosciences with spatial cloud computing, by C. Yang, Y. Xu, and D. Nebert, 2013, International Journal of Digital Earth, 6:4. 297-312.

[i] This paper was approved the NGAC Landsat Advisory Group on December 6, 2013 and adopted by the NGAC as whole on December 11, 2013. The members of the Landsat Advisory Group are: Kass Green, Kass Green & Associates (Co-Chair); Roger Mitchell, MDA Information Systems, Inc. (Co-Chair); Peter Becker (ESRI); John Copple, Sanborn Map Co.; David Cowen, Univ. of South Carolina; Joanne Irene Gabrynowicz, Univ. of Mississippi; Rebecca Moore, Google, Inc.; Tony Spicci, State of Missouri; Cory Springer, Ball Aerospace & Technologies Corp.; Darrel Williams, Global Science & Technology, Inc.; Tony Willardson, Western States Water Council.

[ii] Federal Geographic Data Committee, Initial 2013 Guidance to the National Geospatial Advisory Committee, March 2013 (http://www.fgdc.gov/ngac/meetings/april-2013/2013-fgdc-guidance-to-ngac.pdf)