Landsat in the Cloud

WGISS-49

Kristi Kline, Project Manager
USGS EROS
kkline@usgs.gov
Agenda

• Infrastructure and Interfaces
• Storage
• Data Management
• Collection Processing
• Distribution and Access
Landsat in the Cloud Update

- Began planning in FY17 for cloud transition with goals to Modernize Processing, Access, and Distribution of Landsat Data
  - Change from a primary business model of downloads to enabling access to the full archive
  - Enable users to interact with the data in an integrated environment
  - Ensure provenance and data stewardship
Milestones & Schedule

- Project Kick-Off: September 28, 2017
- Systems Requirements Review/Preliminary Design Review: April 25, 2019
- Critical Design Review: August 15, 2019
- Test Readiness Review: April 17, 2020
- Operational Readiness Review: CY Q2 2020
- Release to Public: CY Q3 2020
Infrastructure and Interfaces
Overall Cloud Environment
Storage
usgs-landsat Bucket Layout

Collection 1
- Level 1: Standard
  - OLI-TIRS
  - ETM
  - TM
  - MSS
- Level 2: Albers
  - OLI-TIRS
  - ETM
  - TM

Collection 2
- Level 1: Standard
  - OLI-TIRS
  - ETM
  - TM
  - MSS
- Level 2: Albers
  - OLI-TIRS
  - ETM
  - TM
Validation and Verification

• To assure the data integrity, a SHA-512 is generated for all files prior to transfer
• When transfer is complete, an AWS Lambda process generates a “Cloud” SHA-512
• Local database tracks both local and cloud-generated SHA values for verification
• Values are also written as metadata on all objects in the cloud for future internal and external transfer validation
• Any discrepancies are double-checked before re-submitting the transfer
• One SHA on-premises takes ~57 seconds
Data Management
Cloud Optimized GeoTIFF Format

- Conducted trade study on cloud formats, which resulted in the selection of Cloud Optimized GeoTIFFs (COGs)
- An enhanced GeoTIFF with tiling and overviews
  - Uses internal tiling instead of lines to speed access and support better remote reading
  - Downsampling is used when lower resolution data is acceptable
  - No changes to the underlying pixels
  - Stored in an unbundled format
  - Data is internally compressed
  - Enables HTTP Get Range requests
- Using Rasterio (GDAL library wrapper)
  - Setting Block Sizes - GeoTIFF = 256; Overview = 128
  - Creating Overviews (2, 4, 8, 16, 32, 64)
  - Compression – Internal, Deflate
SpatioTemporal Asset Catalog (STAC)

• New collaborative standard for managing access metadata
  • Open-source, headed by Planet Labs, freely available on GitHub, working on Landsat extension
  • Flexibility to support many types of geospatial data (satellite, drone, radar, etc.)
  • Allows for interoperability between satellite metadata (e.g. Landsat 8 + Sentinel 2)
  • Lives alongside product-level metadata (MTL, XML)

• Exposes data in a common, machine-readable JSON format for both end users and internal processes

• Includes direct links to S3 objects
• Can be exploited through Jupyter Notebooks by end users to read data directly from the cloud without downloading

• Gaining wide adoption by the remote sensing community
  • i.e., Government, International, Commercial, Academic
Collection Processing
Cloud Image Processing System Design

1. Staging
   • Temp area for input scene list

2. Management Instance
   • Populates db with input scenes
   • Job scanner provides job status
   • Interface to OUI for Batch processing

3. Batch Processing
   • Manage jobs running on EC2

4. CloudWatch
   • Logging and system metrics

5. S3
   • Output products, logs, characterization, inventory metadata

6. Logs
   • CloudWatch and archived to S3

7. PostgreSQL
   • Holds the input scene list
   • Records job status

8. Operator User Interface

9. API Gateway
   • Interface to Lambda functions

10. Lambda Functions
   • Triggering of job scanner
Processing Status Updates

1. Submission process
   - Updates status to Submitted

2. Batch worker
   - Updates status at end of job
   - Success and Failed

3. Job scanner
   - Use cases where worker script is not able to update (loss of instance)
   - Updates reason for failure
   - Creates failure files for job array runs
Distribution and Access
Cloud Smart Design w/On-Premises Systems

- **EarthExplorer (EE)**
  - EE data delivery through Cloud will enable band subsetting to users
  - The EROS Registration System (ERS) will work with Cognito to authenticate users accessing data in the cloud
  - Download metrics will be captured and integrated with the metrics system

- **Machine-2-Machine (M2M)**
  - The M2M API is a RESTful JSON-based Service that provides data access, distribution and ordering
  - All requests are logged and provide a standard response that can be scripted by users to implement their own data discovery and download based on their needs
Cloud Integration with EarthExplorer

- EarthExplorer will be configured to point to cloud data store locations
  - Familiar interface
  - A new download option will be displayed
  - The download option will take users to a STAC Browser Page, where users will be able to select individual bands
Single Band Download from EarthExplorer

Download Options

- Download All Bands
- Select Bands

Level-2 Surface Reflectance Cloud-Optimized GeoTIFF
Level-2 Surface Temperature Cloud-Optimized GeoTIFF

Notes on Download Options:
In the rare case the Level-2 product has been cleaned up and not regenerated yet, the download will be grayed out -- back in EE's Results tab, the item Backot will be available.

- LE07_L2SP_040033_19990929_20190822_02_T1_ST_ATRAN.TIF
- LE07_L2SP_040033_19990929_20190822_02_T1_ST_B6_VCID_1.TIF
- LE07_L2SP_040033_19990929_20190822_02_T1_ST_CDIST.TIF
- etc....
- LE07_L2SP_040033_19990929_20190822_02_T1_MTL.txt
- LE07_L2SP_040033_19990929_20190822_02_T1_MTL.xml

[Buttons]: Deselect All, Select All, Download Selected, Cancel
Sat-api Browser Demo

- [http://sat-api-browser.s3-website-us-east-1.amazonaws.com/](http://sat-api-browser.s3-website-us-east-1.amazonaws.com/)
New LandsatLook Tool
## Pangeo

### Interchangeable Pieces in Pangeo (Pick 1 or More From Each Row)

<table>
<thead>
<tr>
<th>Data Models</th>
<th>xarray</th>
<th>Iris</th>
<th>pandas</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-D Arrays</td>
<td>NumPy</td>
<td>DASK</td>
<td></td>
</tr>
<tr>
<td>Processing Mode</td>
<td>Interactive</td>
<td>Batch</td>
<td>Serverless</td>
</tr>
<tr>
<td>Compute Platform</td>
<td>HPC</td>
<td>AWS</td>
<td>Google Cloud Platform</td>
</tr>
<tr>
<td>Foundation</td>
<td>Python</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Egress Filter Design Details – Limiting Costs

- **Design limits users based on system threshold parameter**
  - Starting value determined – PB/time period
  - Value is stored in Secrets Manager
  - WonderShaper is a Linux tool that limits bandwidth per EC2
  - Adjust any time during a time period

- **Auto-scaling cluster with Nginx and WonderShaper configured to limit bandwidth**
  - Nginx provides limiting by IP and total number of connections
  - WonderShaper will monitor and limit network bandwidth
    - Max Kb/s download rate is configurable based on the System_threshold parameter defined by the USGS
Egress Use Cases

1. Traditional User
   A. Occasional user
      i. Uses EarthExplorer or LandsatLook
      ii. Non-regular interaction
      iii. Lower volume
      iv. Rate limited
   B. Persistent User
      i. Some scripting using M2M
      ii. Using scripting or tools on a more frequent basis
      iii. Higher volume
      iv. Rate limited

2. Non US-Oregon-West User
   A. AWS account outside US-Oregon-West
   B. Higher volume
   C. Rate limited

3. Direct Access User (Oregon West)
   A. AWS Account in Oregon West
   B. Higher volume
   C. Direct read access to the data
   D. NOT rate limited

4. Rogue Actor
   A. DDoS – Stopped from gaining access
   B. Logged for further action