Landsat Authenticity and Integrity of Replicated Data

March 23, 2022

USGS

U.S. Department of the Interior
U.S. Geological Survey

Summary

- Ensuring data integrity and authenticity of Landsat data is of utmost importance
- USGS enabled Landsat in the Cloud in 2020 and utilizes a hybrid cloud approach for data processing, storage and access
- Checksums are currently utilized by internal Landsat processes to ensure data integrity
- Landsat users retrieving data from the USGS cloud archive are provided a checksum value
- USGS will be investigating strategies and technologies to specify and verify replica archive copies as well as individual products for data security, data integrity, digital provenance, and authenticity.

Landsat Data Integrity

- **Definitions related to data**
  - Authoritative Data Source – the datastore or system that contains and provides the data and products that are considered to be the primary source for this information
  - Data Authenticity – ensures the digital provenance of the digital object from creation through its entire lifetime
  - Data Integrity – protecting data from unauthorized changes

Landsat Data Integrity

- **Landsat data is provided and replicated in multiple locations**
- **Ensuring authenticity and integrity of Landsat data in each location is of utmost importance to both Landsat and customers**
- **Checksum - a sum derived from the bits of a segment of computer data that is calculated before and after transmission or storage to assure that the data is free from errors or tampering**
- **Checksums are utilized by Landsat software in multiple places to ensure data integrity is maintained as it is transmitted and stored from one place to another**
  - Multiple locations on-premises at EROS
  - In the cloud environment

Landsat Data Integrity (continued)

- **Checksum types**
  - Message Digest Algorithm (MD5)
    - Currently used on-premises
    - 128-bit has value expressed in text formal as 32-digit hexadecimal
  - Secure Hash Algorithm (SHA)-512 – used in the cloud environment
    - Currently used in the cloud
    - Represented by 128 characters in hex format
    - *preferred method recommended by NIST
- **Eventually Landsat will convert all checksum checks to use SHA-512**
- **At each step in the transmission and storage of Landsat data a checksum is performed to ensure data integrity from one location to the next**
- **Customers who pull Landsat data, say from an S3 bucket (storage) location in the cloud are provided a checksum value with which to compare on their side once the data is retrieved, thus ensuring data integrity on the customer side**

Other Options USGS is Exploring

- **Data Hash in Metadata**
- **Advantages:**
  - Included with the data, no need to find and access a separate resource
  - Encourage development of approaches/tool for validating exact matching for input Landsat data

Other Options USGS is Exploring (continued)

- **Non-Fungible Token (NFT)**
  - Popularized by Ethereum community several years ago in EIP-721 (Ethereum Improvement Proposal)
- **Advantages:**
  - Widely applicable to a broad universe of distinguishable digital assets
  - Can be used to manage various forms of entitlement
    - Authority to create revisions
    - Other processes that rely on the answer to the question: "who owns this thing?" in a way that supports transfer of ownership
    - Blockchain + IPFS + IPLD seem sufficient to cover the data integrity, data authenticity, provenance, traced modifications problems, **and distribution**