Cloud-Optimized Tools for the Surface Biology & Geology High-Frequency Time Series Campaign

Johana Chazaro (California Baptist University),
Marina Dunn (UC Riverside),
Dieu My Nguyen (CU Boulder)

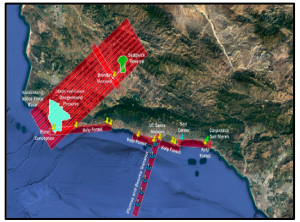
Mentor: Dr. Alexey Shiklomanov (GSFC-618)
Category: Science, Computer Science
May 10, 2022 – 2-4pm EDT: Teams Link

Abstract

The Surface Biology and Geology High-Frequency Time Series (SHIFT) campaign's scientists required a cloud-optimized approach to accessing hyperspectral imaging data of vegetation. Our team proposed and designed a data pipeline that transforms the raw spectral data into Zarr format. This pipeline provides an organized and simplified approach to extracting, transforming, visualizing and analyzing data through a Science Managed Cloud Environment (SMCE).

Introduction and Background

- The SBG-SHIFT campaign hosted by JPL utilizes the Airborne Visible InfraRed Imaging Spectrometer (AVIRIS) to collect spectral data of land and aquatic plant communities over an area in Southern California.
- Our team tests and develops cloud-based tools for data storage, processing, and analysis. These efficient methods will help scientists track the health and resilience of plant communities in the region.



Methods

Resources:

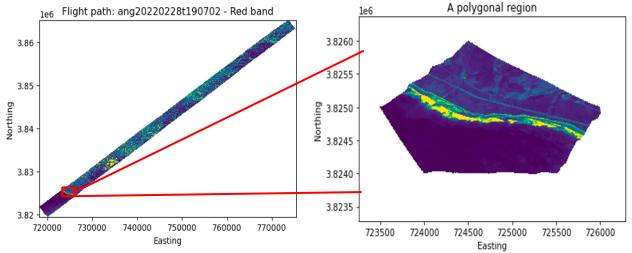
- SMCE Jupyter Hub
- NCCS HPC cluster
- AWS S3, CLI
- AVIRIS-NG data from JPL portal

Create data intake and transformation pipeline:

1. Retrieve raw data from JPL Portal
2. Develop Python code transforming L1/L2 data (geometry, reflectance, radiance) to Zarr
3. Store data on AWS S3 in a STAC-compliant format
4. Create sample data analysis and visualizations

Results

- Developed and documented code on Github.
- Created interactive notebooks demonstrating accessing data on S3, performing common data operations, and visualizing.



Conclusions

We developed a pipeline to efficiently transform and store SHIFT data in an analysis-ready, cloud-optimized (ARCO) and SpatioTemporal Asset Catalog (STAC) specification-compliant format. This will allow for easier indexing, discovery, and analysis. We provide detailed documentation of the pipeline for future users of this AVIRIS data.

Challenges faced:

- SMCE home storage was not persistent
  - *Solution*: notebooks and files used to run data analysis were stored in an S3 bucket
- SMCE alone didn't have the computing capabilities to support Zarr archive creation
  - *Solution*: use NCCS Discover HPC cluster instead

Future Goals/Considerations:

- Generate Zarr archives without requiring HPC.
- Compare performance of different data types (e.g., Zarr vs. GeoTIFF).
- Apply methods to additional Earth science dataset for visualization & analysis

References & Acknowledgements

NAVTECA, JPL AVIRIS-NG Team, SMCE, DevelopmentSeed
https://avng.jpl.nasa.gov/pub/SHIFT/v0/
https://github.com/marinadunn/SHIFT-STAC-backend