Final Report

- Introduction
- User Experience
- DAAC Collaboration
- Product Status
    - Analysis Ready Data
        - GDAL Community Engagement
    - Transformation Workflows
    - Sentinel-2
    - Terra Fusion
    - Terra Fusion vs. Sentinel-2
- Recommendation
    - Data Transformation
    - Web Mapping Service
    - Web Coverage Services
- Future Work
    - GDAL
    - ArcGIS
    - Harmony and Giovanni

Introduction

We've developed a systematic data transformation methodology to enable Web Coverage Services (WCS) and ArcGIS Image Services within ESDIS Cumulus Cloud to increase usability/interoperability of data in commonly used geospatial tools. Many NASA Earthdata products are distributed in scientific data archiving formats like netCDF and HDF. However, such data may not be easily consumed by end users who use GIS tools that can't open or visualize data properly on maps. Our project goal is to narrow the gap by rapidly enabling image services that can quickly feed scientific data in images to GIS users through the use of cloud technology. For cloud technology, we focused on the use of EC2 instances for image services, API Gateways for user facing interface, AWS Lambda functions for data transformation, and S3 object storage system for data input and output because they are core parts of ESDIS CUMULUS.

This document points to How-to articles that point to specific code examples that are developed for last 2 years. If you find any issue in our document or software, please submit a ticket.

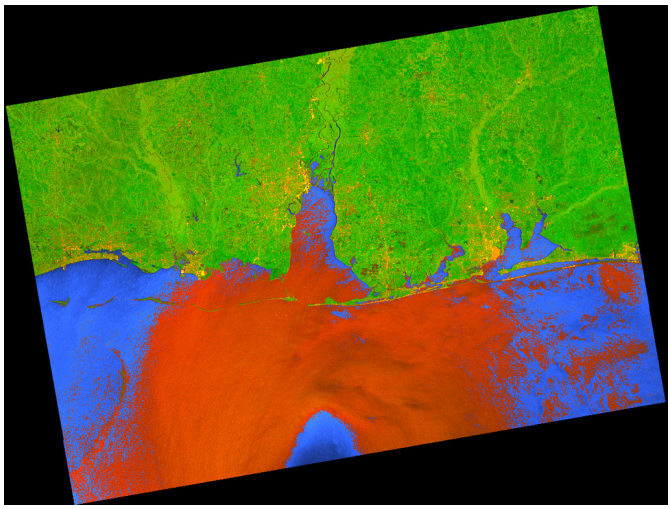
This document will provide insight for both DAACs and end users who want to transform NASA Earthdata formats (e.g., HDF, NetCDF) into different analysis-ready data (ARD) formats and geospatial data services using technologies that can effectively run in the Amazon Web Services (AWS) cloud computing environment.

User Experience

The successful user experience for NASA Earthdata is about meeting a data consumer's needs on individual level. For example, mobile app users will look for data that will fit in their tiny memory and space with streaming service. Alexa users will look for data can be delivered by voice. Cloud users will look for data that can be analyzed easily with commercial-off-the-shelf (COTS) solutions provided by cloud service vendor. Therefore, systematic data transformation of the existing data is necessary to enhance the user's experience and transforming scientific data into image data is one possibility to enhance user's experience.

Why image? Image is the format that everyone, including AI/ML, can enjoy! That's the main reason why NASA Earth Observatory publishes the image of day. You can engage not only general public but also students who can become future citizen Earth scientists through gamification of NASA Earthdata.

Hurricane Sally Image from ASF	Minecraft Map from Hurricane Sally Image



Why ArcGIS? Esri's ArcGIS Enterprise is an amazing all-inclusive turn-key solution for NASA Earthdata in both cloud and on-premise systems. It can cover the most basic services that EOSDIS provides and can go beyond to maximize user experience.

- Data Search to Services in 0 seconds: Services are mostly empty other than downloading data in Earthdata search. ArcGIS Portal links data to services automatically.
- Image of the Day to Story Map with live data visualization: Story Map can give a life to data. It can turn a museum into a zoo.
- Web Application for Mobile users
- Jupyter Notebooks for sharing code and data
- Dashboard for combining metrics from all kinds of data sources (e.g., COVID-19 / SEDAC data): An alternative way to summarize and visualize data, particularly for Swath.

DAAC Collaboration

DAACs play the critical role in interacting with real data users and they care most about UX.

ASDC infused our technology and created several image services in on-prem systems.

Although our focus was ASDC products, our workflows can be easily extended to other DAACs. We successfully collaborated with ASF and GSFC.

Product Status

Analysis Ready Data

A good UX hinges on the availability of Analysis-Ready-Data (ARD) that can quickly deliver data that user needs. We measured the performance of ARD candidates and found out that CRF can deliver the best UX in ArcGIS. Therefore, we have developed several transformation workflows that are built around the latest GDAL.

GDAL Community Engagement

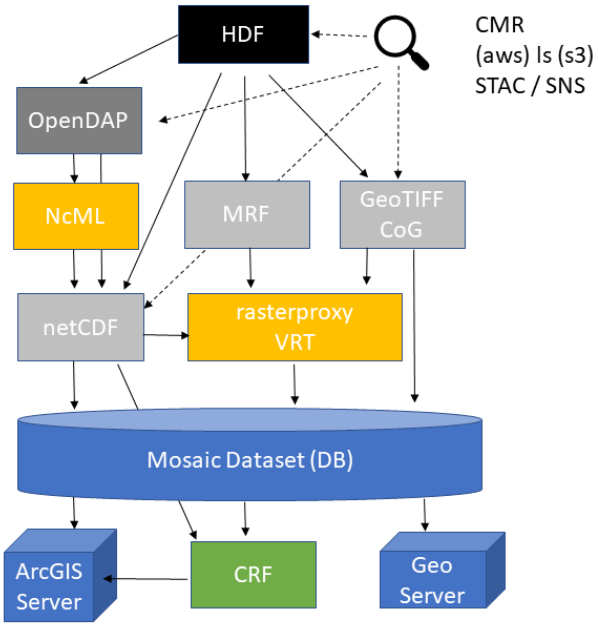
Through BEDI, GDAL Enhancement for ESDIS (GEE) project identified an issue in handling multi-dimensional dataset and made patches. When the GEE team made a pull request after clearing NOSA, GDAL community reviewed the requests and created a new RFC 75 to generalize our patch work further. The RFC was discussed, approved, and fully implemented in GDAL 3.1.

The large GDAL community now can easily access and transform arbitrary N-dimensional dataset that can be found in netCDF and HDF. The new GDAL 3.1 multidimensional APIs and tools also supports group hierarchy so users can unambiguously extract and subset data.

Any NASA HDF data product that has transposed X and Y dimension can benefit from the new GDAL 3.1 capability. This new capability is already tested through SDT project and proven to work in cloud as well. For example, creating GeoTIFF image from MOPITT 5D dataset in a large (~30G) TERRA FUSION granule on AWS S3 was possible through the new gdalmdimtranslate command line tool.

Transformation Workflows

We investigated the most commonly used workflows that can be applied to NASA CUMULUS. The workflows are illustrated in the diagram as arrows. We've developed 20 AWS Lambda functions that correspond to the arrows. When they are put together and run in steps, they can enable image services for both ArcGIS Server and GeoServer. They can be easily deployed via Serverless Framework.



In the above diagram, please note that you can create mosaic dataset or CRF directly from an aggregated, CF-compliant netCDF along time dimension.

CRF, at the bottom of the diagram, needs a special attention. The Cloud Raster Format is an Esri-created raster format that is optimized for writing and reading large files in a distributed processing and storage environment. We ran a few experiments to measure the performance of CRF and the performance of CRF is qutie amazing. There's only 0.4 second difference when CRF is put on S3 instead of local drive. This is quite remarkable because both THREDDS (e.g., Terra Fusion 600X slower Local vs. S3) and Hyrax (e.g., 90X slower EFS vs. S3) performs very poorly when data are served from S3.

Sentinel-2

Recently, we learned that Sentinel-2 data are available from AWS Open Data Registry. We wanted to know if we can replicate what Esri did with Sentinel-2. We tested CoG files for Aerosol Optical Thickness (AOT) with Lambda functions that we've developed for this project. It took only a day's effort to create an image service out of Sentinel-2 CoG data. This experiment validates that our approach can be easily adapted by NASA DAACs and end-users. It would be great if NASA Earthdata provides an authoritative ArcGIS Image Service, not Esri.

Terra Fusion

Terra Fusion is a NASA ACCESS 2015 project. Terra Fusion is an ultimate test dataset for the existing software for netCDF on cloud because file size is huge. Terra Fusion helped us to find issues in several open source projects like Hyrax, THREDDS, and GDAL. For example, GDAL alone can't handle Terra Fusion properly since netCDF swath handling needs improvement although GDAL can read data from S3 efficiently. SDT is necessary to read lat/lon and reproject swath to grid. We subsetted MODIS and created an aggregated netCDF. Then, we created a mosaic dataset directly from the aggregated netCDF. There are a lot of interesting technical details but the valuable lesson is that the meeting the CF conventions alone is not enough to make dataset fully interoperable with the current GDAL and ArcGIS.

Terra Fusion vs. Sentinel-2

It's worth nothing the two extreme approaches that Terra Fusion and Sentinel-2 took for cloud and compare them. What's convenient for atmospheric data scientists in super-computing environment may not work well for general public in cloud-computing environment. It's time to update NASA Data Producer's guide and Data-Interoperability-Working Group (DIWG) recommendations to addresses the issues when data are put into cloud. Data usage should drive the final delivery format on cloud. Sentinel-2's approach seems better than Terra Fusion but we believe that CRF would be more usable than CoG as analysis ready data format in cloud.

Recommendation

Data Transformation

Data transformation can be done independently without Esri ArcGIS software such as ArcPy. Using the latest GDAL python or CLI is recommended for transformation if ArcPy lacks what the latest GDAL can provide.

Some ArcPy functions need Portal Signin capability. Therefore, installing ArcGIS Portal and federating it with ArcGIS Server is recommended.

In AWS environment, web proxy installation can be skipped and native AWS web service front-end can be used.

If input data source is not on S3 through CUMULUS, consider using OPeNDAP to subset data. Use NcML to modify some attributes and overwrite variable to make data CF-compliant.

If ArcGIS Pro on Windows is used for creating mosaic dataset, working with a large input table is fine.

If ArcPy and MDCS on Linux Windows Server is used, there's a limit (99 rows) in processing input table for creating mosaic dataset. Input source doesn't matter. Neither CSV nor netCDF works.

If MDCS fails, export the input table into CSV and modify header to standardize the key field names such as StdZ, StdTime. Then, try ArcGIS Pro to create a large mosaic dataset using the CSV file.

If AWS Lambda will be used for data transformation, make sure that data size is small. Data transfer & transformation should not take longer than 15 minutes.

Creating input table on AWS RDS is recommended for parallel / asynchronous data transformation through Lambda. Make sure that table has unique key to avoid storing duplicate entries.

Convert any mosaic dataset into CRF to optimize service performance especially on cloud. It will cost extra for storage but use it to improve user experience.

MDCS can publish service to Server in a specific folder but cannot to Portal in a specific folder. Use ArcPy to move the service to a specific folder in Portal.

Build CRF and serve it via ArcGIS Server for popular dataset. Although building CRF may take extra time and CRF takes up extra storage, it's worth creating for the best user experience.

Web Mapping Service

- When using grouped WMS services it is recommended to view an individual sublayer of the WMS to improve performance and minimize requests being sent to the WMS server.
- When viewing multiple WMS services it is recommended to turn off legend for layers not being used to minimize server requests in thelayer properties options.
- WMS should be used primarily for visual representation and rendering of data on a map.
- WMS has limited capability when needing to run any analytical processes such as querying, extracting, or binning. Selecting points of interest and popup configuration is not supported with WMS services.
- WMS layers can support styles which can be used as custom parameters for layer requests. Custom parameters are applied to all layers that you specify to be added to the map. It is recommended that if user wants to append different custom parameters to specific layers only, user must add that layer separately.
- It is recommended when configuring the image service to enable caching on demand so that users can navigate to areas of interest which will then generate cached tiles and increase performance and image quality of the WMS service.
- If WMS service allows clients to cache tiles locally it is recommended that the clients must manually clear their cache to get any newer tiles from the server. If you will be updating your cache frequently it is best not to use this option.
- Identifying features will only work on features from the topmost layer in the WMS services drawing hierarchy, it is also recommended to set your identify tolerance levels based on user requirements. This can be done when configuring the service on the WMS server
- OGC WMS coordinate system needs to be the same coordinate system for the basemap in order to display correctly. When creating the image service it is recommended to use the coordinate system required by users as they are unable to change the coordinate system of the WMS

Web Coverage Services

- To configure your WCS service to use external capabilities files, it is recommended to update the WCS service to reference the URL of your capabilities files. You can do this in ArcGIS Server Manager
- WCS services are useful if you want to make your raster datasets available online in an open, recognized way across different platforms and clients. Recommend that client be built to support the WCS 1.0.0, 1.1.0, 1.1.1, 1.1.2, and 2.0.1 specifications can view and work with your service.
- If using multidimensional data in WCS 2.0.1, it's recommended that you create a mosaic dataset using corresponding raster types (for example, netCDF, HDF, and GRIB). Variables are mapped to rangeType in the coverage description and accessed through the range subsetting.
- Dimensions (for example, x, y, t, and z) are mapped to domainSet and accessed through the dimension trimming and slicing.
- Due to the XML schema, time is represented using doubles by converting the UTC date to OLE Automation Date.
- WCS services that are expected to be accessed via OGC interfaces should be secured using Integrated Windows Authentication, HTTP Basic, or HTTP Digest. Most OGC clients (both non-Esri and Esri clients) will understand and work with these widespread standard authentication schemes.
- It is recommended that querying parameters be avoided in the connection url. A WCS capabilities URL is not a valid input. Query parameters should be carried on in GetCapabilities, DescribeCoverage, and GetCoverage requests.

Future Work

The following sections list new features that are desirable in near future. Cloud technology has disrupted many traditional technologies and closed the doors of many high-tech companies (e.g., MapR, Hortonworks). This trend will only accelerate as cloud vendors provide more COTS and Data as a Service. For example, do we still need OPeNDAP to subset and deliver data over network while S3 can do it faster and in much larger scale? We will provide some insights for cloud-era.

GDAL

- Add input file option for gdalmdimtranslate so that netCDF-4 compatible file with .h5 extension can be opened by netCDF driver instead of HDF5 driver.
- Improve Swath handling in HDF/netCDF drivers.
- Fix any issue that doesn't work with the existing NASA Earthdata products.

ArcGIS

- Support STAC.
- Support more Raster Types as a source in Raster Proxy:
  - OPeNDAP URL: Esri was reluctant to add it due to performance, reliability, and security concern.
  - DMR++
  - Zarr
  - Harmony API
- Remove errors that only occur on Wine in Linux but don't appear on ArcGIS Pro in Windows Server.
- Adjust time slider on Portal image layer viewer for "Has Live" data when new data arrives. User needs to configure time slider setting manually to see a new data.

Harmony and Giovanni

- Provide CRF as an output option when data aggregation request is made.
- Avoid duplicate efforts - use ArcGIS Server as back-end for any service that ArcGIS Server already provides.
- Expand Harmony API to invoke raster analytics functions through ArcGIS Server.

File	Modified
Multimedia File harmony_ceres_image_service.mp4	Oct 12, 2020 by Hyokyung Lee
Multimedia File arcgis_pro_multi_dims.mp4	Oct 12, 2020 by Hyokyung Lee
Multimedia File arcgis_pro_no_dims.mp4	Oct 12, 2020 by Hyokyung Lee
Multimedia File aws_lambda.mp4	Oct 12, 2020 by Hyokyung Lee
Multimedia File aws_s3_bucket.mp4	Oct 12, 2020 by Hyokyung Lee
Multimedia File aws_s3_cp.mp4	Oct 12, 2020 by Hyokyung Lee
Multimedia File arcgis_pro_md_input.mp4	Oct 12, 2020 by Hyokyung Lee
Multimedia File mop03j2md_lambda.mp4	Oct 12, 2020 by Hyokyung Lee
Multimedia File mop03j2mrf_lambda.mp4	Oct 12, 2020 by Hyokyung Lee
Multimedia File portal_item_creation.mp4	Oct 12, 2020 by Hyokyung Lee
Multimedia File portal_item_new_map.mp4	Oct 12, 2020 by Hyokyung Lee
Multimedia File portal_item_update.mp4	Oct 12, 2020 by Hyokyung Lee
Multimedia File serverless.mp4	Oct 12, 2020 by Hyokyung Lee
JPEG File S1B_IW_GRDH_1SDV_20200915T235327_20200915T235352_023393_02C6F9_D515.jpg	Oct 15, 2020 by Hyokyung Lee
PNG File sally_minecraft.png	Oct 15, 2020 by Hyokyung Lee
PNG File workflows.png	Oct 16, 2020 by Hyokyung Lee

Download All