Project Details

Name of project: Cloud based analytics for TerraFusion data

Project lead and contact details: John Readey, jreadey@hdfgroup.org, 425-999-6210

Project partners and contact details: Aleksandar Jelenak mailto:ajelenak@hdfgroup ; Kent Yang mailto:yang6@hdfgroup.org

Proposed start and end date: Feb 1, 2020 through Aug 1, 2020

Budget Requested: $5,000

Budget Summary: 50 hours of development work at $100/hour

Project Outline

Project description:

Terra Fusion (https://earthdata.nasa.gov/esds/competitive-programs/access/terra-data-fusion-products) is a NASA ACCESS project that had the goal of combining data from different Terra Earth Observing satellite instruments into a cohesive data product. By combining these data products, it makes it much easier for researchers to use the full spectrum of collected data to understand the dynamics of Earth and its atmosphere over the course of the Terra mission. The Terra Fusion team produced 2.4 PB of HDF5/NetCDF4 data files that are now available for researchers.

There are three challenges to utilizing this data resource that this project hopes to address: efficient access in the cloud, the lack of Python-based tools and examples, and a community-based data platform.

First, for efficient data access, since the size of the collection is so large (even individual files are in the range of 30-50 GB), it is impractical for researchers to download them for local processing. What is needed is the capability of in-situ processing, i.e. bringing the analytics to where the data resides. This was made more feasible by the recent inclusion of the Terra Fusion data in Amazon's Open Data on AWS (https://registry.opendata.aws/terrafusion/). By providing the data in an AWS S3 Public Bucket, codes running on AWS have high throughput/low latency access to this data. Until recently though, this still meant that files had to be copied from S3 to a POSIX file system before they could be accessed with tools typically used by researchers. However recent work by The HDF Group and others have added new capabilities that allow HDF5 data to be accessed directly from S3. For example, the latest release of the HDF5 library provides a plugin (the S3VFD) that enables reading directly from HDF5/NetCDF4 files in S3. Another tool of interest is a REST Service for HDF5, HSDS. This service was developed under another NASA ACCESS project (https://earthdata.nasa.gov/esds/competitive-programs/access/hsds) whose PI is the author of this proposal. HSDS enables server-side parallelism to greater accelerate common data access patterns.

The second challenge with utilizing the Terra Fusion dataset is that lack of Python-based tooling, and examples. Python and associated packages such as xarray have become very popular with the Earth Science community. The Terra Fusion team has created C++ codes for data resampling (see: https://github.com/TerraFusion/advancedFusion), but has few codes in Python that illustrate how the data can be accessed and visualized. This project would develop Python utility libraries, sample Python notebooks, and tutorials that would illustrate best practices with working with the data. These codes would work with Terra Fusion data as it resides in S3 using the techniques described above.

Lastly, this project will provide a data platform to facilitate access to the data. While researchers have the ability to directly launch their own machines on the AWS platform, the skill required, complexity, and costs can be daunting. The JupyterHub/JupyterLab projects make running codes in the cloud more accessible, since no AWS account, or special skills are required. Users just sign in from their browsers and have a ready-to-go environment for running codes in the cloud. The HDF Group has provided such an environment, Kita Lab (https://www.hdfgroup.org/hdfkitalab/) that provides a JupyterHub/JupyterLab access for running codes on AWS. Kita Lab provides access to the HSDS server, HDF5 tools, sample data, and sample Notebooks for Earth Science. For this project, we would utilize the Kita Lab environment to provide access to the Terra Fusion data, Python packages, and sample notebooks developed for this project. Access to Kita Lab is free for anyone interested in trying it out. Conveniently, Kita Lab, and the Terra Data S3 bucket are both hosted in the same AWS region, us-west-2. This means that data access will be much superior to what would be the case if cross-region access was needed.

Project objectives, significance, and impact:

This project has the following objectives:

1. Develop Python packages to facilitate access to Terra Fusion data on AWS
2. Provide sample Python notebooks that illustrate how Terra Fusion data can be analyzed
3. Understand the performance bottlenecks with using HDF5 tools and HSDS with Terra Fusion data; and explore how these can be rectified
4. Promote the use of this data resource through blog articles, webinars, ESIP sessions, etc.
5. Educate the community on the most effective was to utilize HDF5 data in the cloud

The impact of this work will be to enable more people to effectively utilize the Terra Fusion data. This in turn will enable more progress to be made on the environmental challenges facing the world.

A further significance will be in using the Terra Fusion on AWS as a case study for accessing Earth Science data in the cloud. NASA is currently planning to transition mission data to AWS. What is unclear is what the best tools and techniques for accessing these data resources will be. Illustrating how The HDF Group tools along with JupyterLab can be used should provide a useful proof point.

Description of key project steps and timeline:

A rough outline is:

May 2020: Experiment with accessing Terra Fusion data
Apr 2020: Develop Python utility packages
May 2020: Create Python Notebook examples, tutorial
June 2020: Benchmarking, performance analysis, refactoring as needed
July 2021: Webinar, blog post, and talks about the project

Description of additional funding currently supporting this work:

The HDF Group has received $12K in AWS Cloud Credits to support Kita Lab

NASA is providing $24K to The HDF Group in 2020 to enhance the HSDS service.

Outreach

What groups/audiences will be engaged in the project?

Terra Fusion developers, any researchers who have interest in the Terra data.

How will you judge the project's impact?

Log in to the Kita Lab site, direct feedback, attendance to webinar or talks.

How will you share the knowledge generated by the project?

All project codes will be hosted on GitHub. Project notebooks will be available on Kita Lab. The HDF Group will use forum posts, web site updates, and email bulletins to keep the community informed about the project. The ESIP Slack channel will also be used to connect with the ESIP community.

Description of *who* (agencies/individuals) should be aware of this project, i.e. potential outreach targets:

- NASA – Kevin Murphy mailto:kevin.j.murphy@nasa.gov EOSDIS lead
- University of Illinois – Larry Girolamo mailto: gdi@illinois.edu, Terra Fusion project lead
- Amazon AWS - Joe Flasher jflasher@amazon.com, Open Geospatial Lead

Project Partners (as applicable)

Description of project partners (agencies/individuals) and their involvement:

The Terra Fusion team will provide consulting as we have questions on Terra Fusion file contents and structure.

How will this project engage members of the ESIP community.

We will reach out through ESIP mailing list and Slack channels as the project is running. We would like to present our finding at the Summer ESIP meeting in 2020.