ADVANCEMENTS IN THE OPEN DATA CUBE AND THE USE OF ANALYSIS READY DATA IN THE CLOUD

[1]Brian Killough, [2]Syed Rizvi, [2]Andrew Lubawy

[1]NASA Langley Research Center, [2]Analytical Mechanics and Associates, Inc.

ABSTRACT

The Open Data Cube (ODC), created and facilitated by the Committee on Earth Observation Satellites (CEOS), is an open source software architecture that continues to gain global popularity through the integration of analysis-ready data (ARD) on cloud computing frameworks. In 2021, CEOS released a new ODC sandbox that provides global users with a free and open programming interface connected to Google Earth Engine datasets. The open source toolset allows users to run application algorithms using a Google Colab Python notebook environment. This tool demonstrates rapid creation of science products anywhere in the world without the need to download and process the satellite data. Basic operation of the tool will support many users but can also be scaled in size and scope to support enhanced user needs. The creation of the ODC sandbox was prompted by the migration of many CEOS ARD satellite datasets to the cloud. The combination of these datasets in an interoperable data cube framework will inspire the creation of many new application products and advance open science.

Index Terms— Open Data Cube, CEOS, Earth Observation, Satellite Data, Analysis Ready Data, Cloud Computing, Sustainable Development Goals

1. OPEN DATA CUBE

To date, the ODC community has facilitated the initiation and advancement of over 70 operational country-level data cubes around the world [1]. Though many of these are in Africa (54 countries using the Digital Earth Africa platform) the number of data cube deployments continues to grow and inquiries continue to be received. In total, the global impact of the ODC is over 110 countries [2].

As the number of deployments has increased, there has been considerable advancement in the core ODC code and associated application algorithms. These code developments are routinely released on GitHub as open source software by various global users. Most of those releases are from Geoscience Australia (GA), the Commonwealth Scientific and Industrial Research Organization (CSIRO), and NASA, but some releases, mostly applications, have come from a few country-based users. In 2020, the ODC initiated an Open Source Geospatial Foundation (OSGeo) project. OSGeo is a non-profit non-governmental organization whose mission is to support and promote the collaborative development of open geospatial technologies and data. It is believed that future development of the ODC, as an OSGeo Project, will enhance code development and support community growth, involvement and interest.

The ultimate vision of the ODC founding partners has always been to achieve a global network of connected regional data cubes that are self-sustaining and share core ODC code and applications algorithms among a vibrant community of users. As we approach the middle of 2021, only Digital Earth Australia [3] and Digital Earth Africa [4] represent the first pieces of this global network. Most recently, there has been work in the Americas and the Pacific Islands that may lead to new regional initiatives connecting all of the countries in that region. It is this ultimate goal that will lead to improved collaboration and sharing of algorithms among global users and continue to spread the ODC brand and its impact.

2. ANALYSIS READY DATA

Analysis-Ready Data (ARD) is the core component of all ODC deployments [5]. Since 2017, the CEOS Land Surface Imaging Virtual Constellation (LSI-VC) team has developed an ARD Framework [6] which includes three elements: product definition, product specifications and product assessment. By the end of 2020, the LSI-VC team had completed mature ARD product specifications for land surface temperature, land surface reflectance and land surface radar backscatter. In addition, there are a number of new product specifications under development including interferometric radar, polarimetric radar, water-leaving aquatic reflectance, nighttime lights surface radiance and LIDAR products.

The development of ARD specifications has produced significant support from global data providers (within CEOS) and commercial industry (non-public datasets). To date, there is only one known example of a CEOS ARD-compliant dataset available to users. This is the Landsat Collection-2

data on Amazon Web Services (AWS). As of late 2020, there are several other datasets under evaluation including Sentinel-2 surface reflectance, available on the Copernicus Hub and AWS, and Sentinel-1 normalized radar backscatter, processed by Sinergise under a contract with NASA and Digital Earth Africa. In addition, there are other sources of processed satellite data (e.g. Google Earth Engine) that may be compliant with ARD specifications, but lack compliance and validation assessments.

The future of ARD is exciting as its production and availability will increase, thereby allowing more interoperable use and enhanced benefit of these satellite time series data products. With the availability of ARD, it is expected that data cubes (e.g. ODC) will be able to exploit these products for improved time series analyses, interoperability between diverse datasets (e.g. combining radar and optical data) and machine learning [7]. In addition, many of the new CEOS ARD specifications refer to radar products which are used by a smaller fraction of the user community, but have great potential for the future. If CEOS is able to promote the routine production and access to radar ARD from Sentinel-1 and ALOS, it will provide a new and exciting resource for many global data cube users and offer many new application products [8].

3. OPEN DATA CUBE SANDBOX

Concept

The Open Data Cube (ODC) Sandbox is a free and open programming interface that connects users to Google Earth Engine datasets. The open source tool allows users to run Python application algorithms using Google's Colab notebook environment. This tool enables the rapid creation of science products without the need to download and process the satellite data. Some example applications include: custom cloud-filtered mosaics, spectral index products including vegetation fractional cover, historic water extent, vegetation phenology and land change. Basic operation of the tool will support many users but can also be scaled in size and scope, with added resources, to support enhanced user needs. The new sandbox can be found at the following web address: *https://www.openearthalliance.org/sandbox*.

Technical Description

The CEOS Systems Engineering Office (SEO) has funded the development of the ODC Sandbox and will manage its operation. Once operational, users can test applications anywhere in the world using Google Earth Engine data. To start, users must have an existing Google account and be an authorized Earth Engine developer. No costs are incurred for these services but prior approval is required and may take several days for Earth Engine approval. The sandbox utilizes a common Google Colab Python notebook environment that provides limited computing (~12 GB RAM) and storage (~70 GB) but with no costs to the user. Though there are some limitations in memory and computing, these resources are sufficient to run analyses over reasonable areas and time windows. In addition, it is expected that the tool can support a large number of simultaneous users for training or testing as the Google Colab environment spawns a dedicated instance for each user opening a sandbox. Also, users can run up to 5 simultaneous sandbox deployments using the same account credentials.

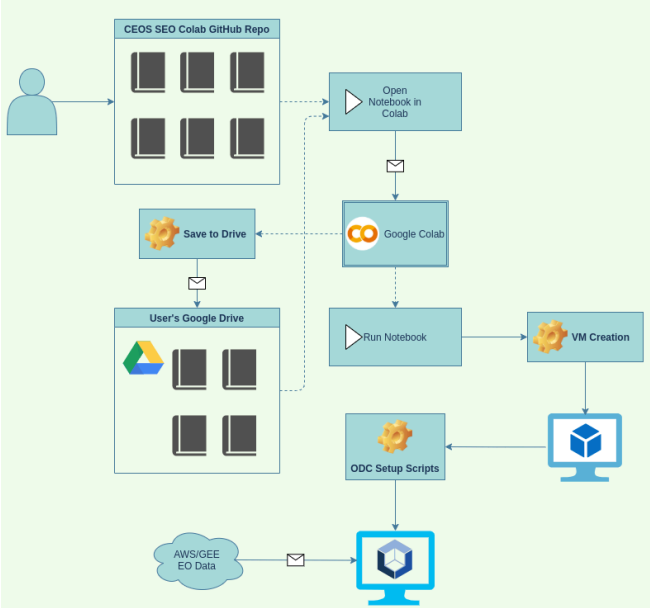

Figure 1. The ODC Colab architecture diagram shows the links between user algorithms (Colab notebooks), satellite datasets, and output products.

Several application notebooks will be available in the baseline sandbox deployment. Each notebook will be prepopulated with code that "runs to completion" and utilizes the satellite data on Earth Engine. Users can make modifications to the code, as desired, in order to run these applications over different locations and times. In addition, documentation and video training resources will be available to support users. All of these open source algorithms and associated resources will be organized and available on GitHub.

Prototype Testing

In early 2020, CEOS, Analytical Mechanics and Associates (AMA), and the Global Partnership for Sustainable Development Data (GPSDD) collaborated with Google to demonstrate the use of the open source Open Data Cube (ODC) technology on the Google Cloud for connecting to the vast supply of Google Earth Engine satellite datasets. It was this collaboration that led to the testing of Google Colab and

the development of the sandbox. This demonstration provides an alternative to the common Google Earth Engine user interface and also an alternative to other cloud computing options with fewer satellite and non-satellite data choices.

The only requirement prior to operation is that the satellite datasets must be "indexed" so that the ODC code can use band name aliases in the analysis notebooks. To date, only Landsat-8, Sentinel-2 and Sentinel-1 data are available in the sandbox environment. Use of other datasets is possible by contacting the SEO for support.

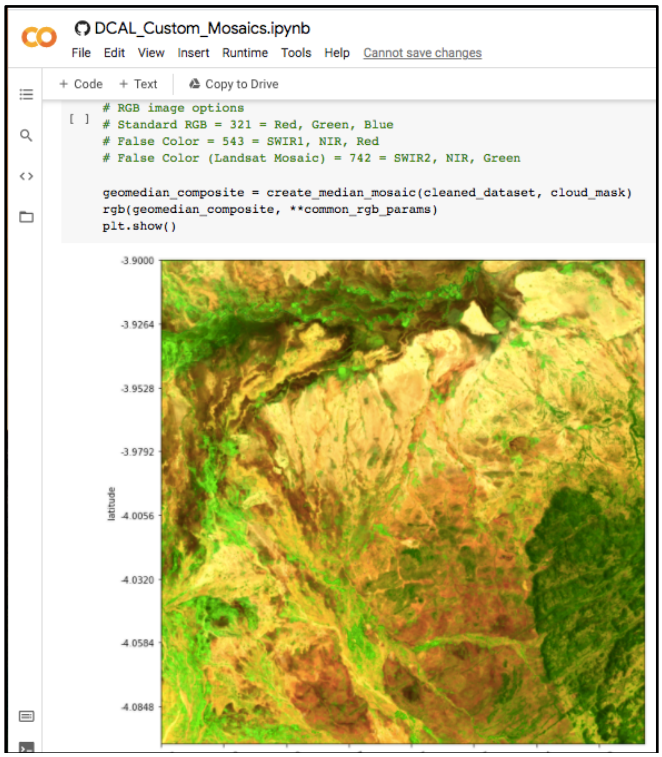

Figure 2. An example Landsat geomedian product from the ODC-Google Sandbox running in the Google Colab environment.

Connections to GEO

The Open Earth Alliance (OEA) is a new GEO Community Activity focused on exploiting satellite data technologies such as the ODC. This new ODC sandbox concept will be a contribution to the OEA initiative as it uses advanced technology to stimulate the use of satellite data for global impact. In addition, this initiative is also aligned with the objectives of the GEO Knowledge Hub to promote "open science" through the use of free/open datasets and shared tools and algorithms. Finally, the ODC sandbox will also include a User Forum to answer user questions and continue to grow the user community. This forum will be similar to the Sentinel Hub User Forum (forum.sentinel-hub.com) and will manage its content using "topics".

4. CONCLUSION

The Open Data Cube (ODC) initiative continues to progress and have impact across the globe. CEOS, as a leader in satellite remote sensing, is facilitating the production and dissemination of ARD and thereby making data easier to access and apply to science applications and decision-making. Through the release of a new ODC sandbox tool, users can now freely access satellite datasets on Google's Earth Engine computing framework. This combination of the ODC and ARD in a sandbox programming environment, will enable global users to address many important problems and determine how satellite data can improve their understanding of the environment and advance open science.

5. REFERENCES

[1] J. Ross, B. Killough, T. Dhu, M.Paget, Open Data Cube and the Committee on Earth Observation Satellites Data Cube Initiative, IAC, 2017.

[2] B. Killough, Overview of the Open Data Cube Initiative, IGARSS, 2018.

[3] D. Gavin, T. Dhu, S. Sagar, N. Mueller, B. Dunn, A. Lewis, et al., Digital Earth Australia - From Satellite Data to Better Decisions, IGARSS, 2018.

[4] B. Killough, The Impact of Analysis Ready Data in the Africa Regional Data Cube, IGARSS, 2019.

[5] B. Killough, A. Siqueira, G. Dyke, Advancements in the Open Data Cube and Analysis Ready Data - Past, Present and Future, IGARSS 2020.

[6] A. Siqueira, A. Lewis, M. Thankappan, Z. Szantoi, B. Killough, P. Goryl, S. Labahn, J. Ross, T. Tadono, A. Rosenqvist, J. Lacey, M. Steventon, CEOS Analysis Ready Data for Land: Implementation Phase and Next Steps, IGARSS, 2020.

[7] Gregory Giuliani, Bruno Chatenoux, Andrea De Bono, Denisa Rodila, Jean-Philippe Richard, Karin Allenbach, Hy Dao & Pascal Peduzzi (2017): Building an Earth Observations Data Cube: lessons learned from the Swiss Data Cube (SDC) on generating Analysis Ready Data (ARD), Big Earth Data, 2017.

[8] A. Rosenqvist, B. Killough, A. Lubawy, J. Rattz. SAR Analysis Tools for the Open Data Cube, IGARSS, 2020.