GENERATING ANALYSIS READY DATA COLLECTIONS FOR BRAZIL

Rennan F. B. Marujo[1] Karine R. Ferreira[1] Gilberto R. Queiroz[1]
Raphael W. Costa[1] Jeferson S. Arcanjo[1] Ricardo C. M. Souza[1]

[1]INPE, National Institute for Space Research, Sao Jose dos Campos, SP, Brazil.
(rennan.marujo, karine.ferreira, gilberto.queiroz, raphael.costa, jeferson.arcanjo,ricardo.cartaxo)@inpe.br

ABSTRACT

Brazil is a country of continental scale area, with a territory of over 8.5 millions of km^2. Nowadays, data from different satellites and sensors with distinct spatial, temporal, and spectral resolutions are available for free. However, prepare and handle these large amounts of data is an exhaustive task. The Brazil Data Cube (BDC) Project emerges in this context processing and preparing Analysis Ready Data (ARD) of medium spatial resolution satellite sensors, condensing it as image data cubes for the Brazilian territory, allowing researchers to easily extract time series from them, focusing on the analysis instead of the processing. This paper describes the workflow to generate BDC ARD input data, which are used as input for data cubes, as well as point our learnings and findings handling and processing these data.

***Index Terms*—** Brazil, Analysis Ready Data, Remote Sensing, Surface Reflectance, Software.

1. INTRODUCTION

Brazil Data Cube (BDC) [1] is being developed by Brazil's National Institute for Space Research (INPE), since January 2019 as part of the *Environmental Monitoring of Brazilian Biomes* project. It creates multidimensional data cubes from medium spatial resolution Analysis Ready Data (ARD), for the Brazilian territory, to generate land use and cover maps through machine learning and time series analysis. Although nowadays we have an unprecedented quantity of available free remote sensing data [2], initiatives to ensure the quality of these data are emerging [3] and innovative initiatives to facilitate the data dissemination are rising [4].

The Brazil Data Cube project has four main objectives: (*i*) create ARD sets from medium spatial resolution remote sensing images, which includes data from the Earth observation satellites CBERS-4, Landsat-8 and Sentinel-2; (*ii*) generate data cubes from these ARD to support image time series analysis; (*iii*) propose, develop and use novel methods and big data technologies to store and process these data cubes; and (*iv*) extract land use and land cover information from the data cubes through image processing procedures, machine learning and time series analysis.

This paper describes the first part of the BDC objective, how BDC processes and prepare ARD of medium spatial, detailing the ingested data, the radiometric processing performed on it and the tools developed to orchestrate these tasks.

2. ARD COLLECTIONS FOR BRAZIL

BDC project uses free medium spatial resolution images from different data providers and processes them to Surface Reflectance (SR) products. The acquisition (download and publishing to database), radiometric corrections and optional processings such as cloud masking and spectral indices calculation is orchestrated by a in-development tool called BDC *Collection Builder*, which is a free and open source software that is being constantly updated to acquire data from new data providers, to handle new types of data and to support new processes. Figure 1 illustrates the processings that are handled by the BDC, which are described in this section.

BDC *Collection Builder* supports acquiring images from Sentinel-2 (A and B), Landsat-5, Landsat-7, Landsat-8, Landsat-9, Terra and Aqua, and can distribute these collections compressed as a single asset or with each file as an individual asset. Images can be saved as Cloud Optimized GeoTIFF (COG), which are regular GeoTIFF files with an internal organization of overviews that enable more efficient access to the data. It can also process these images, e. g. perform an atmosphere correction or cloud masking, and after this, these images and their assets are made available using the Spatio Temporal Asset Catalog (STAC) [4], which is an open specification based on JSON and RESTful to increase interoperability of searching for geospatial data.

2.1. Input images

The main satellite/sensor images that are acquired and processed by BDC are CBERS-4 (MUX and WFI), CBERS-4A

This research was supported by the Amazon Fund through the financial collaboration of the Brazilian Development Bank (BNDES) and the Foundation for Science, Technology and Space Applications (FUNCATE), process 17.2.0536.1.

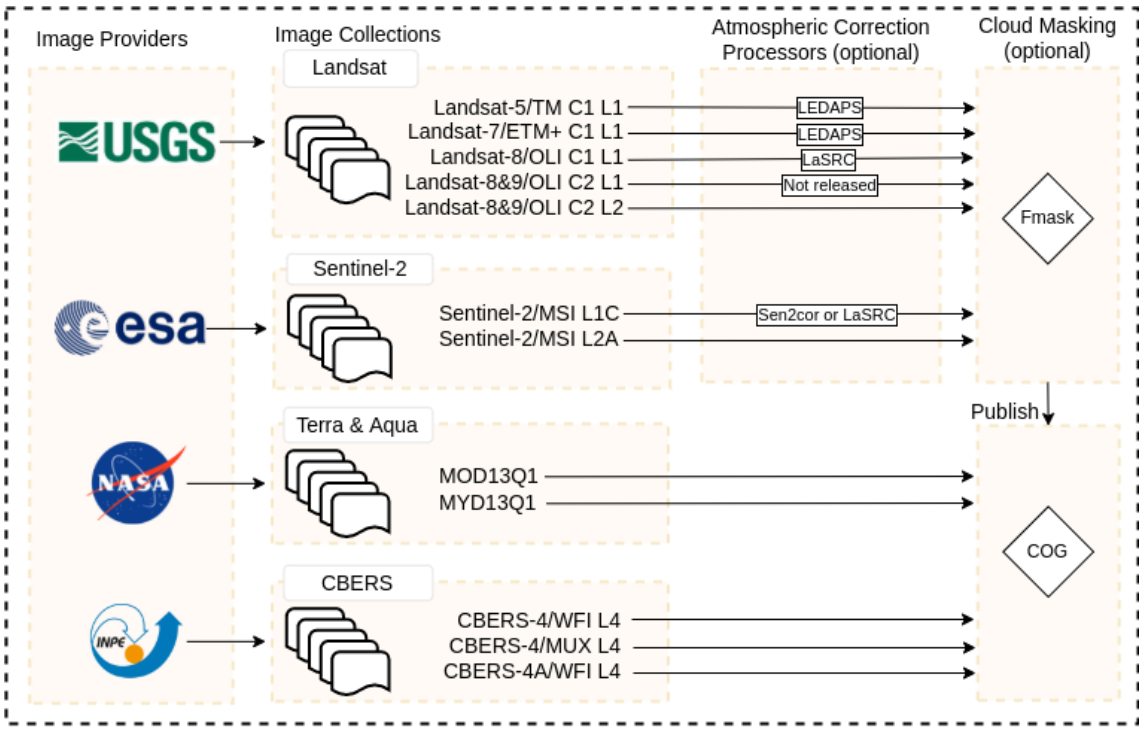

Fig. 1. BDC Collection Builder Workflow.

(WFI), Landsat-8 Operational Land Imager (OLI), Sentinel-2 (2A and 2B) MultiSpectral Instrument (MSI) and from AQUA/TERRA Moderate Resolution Imaging Spectroradiometer (MODIS) sensors, with plans to support Amazonia-1 (WFI).

BDC *collection builder* is capable of acquiring images from the satellites of the China–Brazil Earth Resources Satellite (CBERS) program directly from INPE, in this case CBERS-4 (MUX and WFI) and CBERS-4A (WFI) as Level 4 (L4) products. These products are already corrected for atmosphere effects, processed as SR products, which are orthorectified, radiometrically corrected with system geometric correction refined by control points and with digital terrain elevation model applied by the MS3 software [5].

It also supports acquiring images from the Landsat program, which include images from the recently launched Landsat-9, that is replacing Landsat-7 [6]. Landsat data is maintained as collections, being Collection 2 mainly improvements the data geometric accuracy, digital elevation modeling and radiometric calibration [7]. BDC *collection builder* is capable of acquiring Collection 1 and Collection 2 data. Collection 1 Level 1 data acquired can be processed through the Landsat Surface Reflectance Code (LaSRC) [8]. Although Landsat Collection 2 Level 1 and Level 2 data can be acquired using BDC *Collection Builder* the Level 1 data cannot be processed by it, since the official software to process it is not available yet, due to a migration of repositories in progress [9].

Regarding Sentinel-2, BDC *collection builder* is capable of acquiring Level-1C and Level-2A products. It can process Level-1C products using LaSRC (as performed for Landsat) [10] or the Sen2cor [11], European Space Agency standard processor for Sentinel-2. Considering Sen2cor, BDC *collection builder* supports different versions of the processor, 2.5.5, 2.8.0, 2.9.0 and 2.10.0 [12], since different dates of L1C products can be only processed using specific versions.

Already processed products from the TERRA and AQUA satellites can also be obtained using the BDC *collection builder*. The BDC project acquire and distribute MOD13Q1 and MYD13Q1 products, which consists in 16-day best pixel composition of spectral bands and vegetation indices (NDVI and EVI) at 250 meters of spatial resolution [13].

2.2. Cloud/Cloud Shadow Mask

Besides atmosphere correction, the BDC *collection builder* is being developed to support different Cloud and Cloud Shadow masking algorithms, as an optional processing. These algorithms rely on the characteristic of each sensor. The LaSRC processor provides cloud and cloud shadow masks (QA_pixel band), when applied to Landsat-8. Similarly, Sen2cor when applied to Sentinel-2, can produce the Scene Classification Layer (SCL), which also contains cloud and cloud shadow information. The MS3 software also provides Cloud Mask for CBERS data [5].

Cloud masking is not perfect and other cloud masking al-

gorithms are constantly evolving. Fmask (version 4.3) [14], for instance, is capable of generating labels, as clear land, clear water and snow/ice and showed improved detection in comparison to the standard cloud masking algorithms [15].

2.3. Cloud Optimized Geotiff and Publish

After collecting and processing the images, the last step of the ARD workflow consists in generating COG files and publishing their metadata in a database. The main advantage of COG is that the GeoTIFF files are internally organized in several overview layers and blocks that allows more efficient load and visualization tasks, common on Cloud and Web based environments, not requiring to load the entire images.

Once published, images can be searched using a STAC service, which will perform the search against a database. The STAC provides uniform ways to discover and access geospatial data and is increasing in popularity in the recent years.

2.4. Implementation Details

As illustrated on Figure 1, the workflow consists in a series of chained processings, for instance a full chain of processing of a Landsat-8 image: (*i*) a Landsat-8 C1 L1 image is acquired from the USGS provider, (*ii*) it is processed to SR through the LaSRC atmospheric correction processor, (*iii*) an improved cloud mask is generated to this data using the Fmask Cloud Masking [14], and (*iv*) all resulted images are converted into COG format and alongside the auxiliary files (.txt/.csv) are published into a database. Independent tasks can run in parallel, for instance, multiple images (from different dates or geographic location) can be processed simultaneously. However, a process may require previous processes to be completed, for instance the atmospheric correction of an image will only be performed after the image is fully acquired (downloaded). Due to that BDC *Collection Builder* works with an orchestrator to manage queues.

In this context the implementation of the BDC *Collection Builder* uses RabbitMQ [16] alongside Celery [17] to manage the processing of tasks. Using the Landsat-8 full chain of processing presented on the last paragraph, Celery receives the task of the chain processing, which is passed to RabbitMQ to organize it in its queues, since the first step consists in acquiring the image, RabbitMQ send this task to a worker, in this case controlled by Celery. After the end of this task, the worker forward the result to rabbit again to be inserted into the atmospheric correction queue, and so goes the chain.

3. RESULTS AND CHALLENGES

Since 2019, BDC has been acquiring, for the entire Brazilian territory, Landsat-8 Collection 1 Level 1, MOD13Q1, MYD13Q1, Sentinel-2 Level-1C and CBERS-4 images ranging from 2016 to 2022. Acquired Landsat-8 Collection 1 Level 1 has been processed, using LaSRC and generating Fmask [14] Cloud masks. The obtained Sentinel-2 is nowadays being processed using ESA's default atmospheric correction processor, the Sen2cor [11], while CBERS data has been obtained already processed as surface reflectance products directly from INPE and MODIS products (MOD13Q1 and MYD13Q1) are also obtained as products, being both CBERS and MODIS data converted to COG format. For a few test sites, BDC was also acquiring and processing Landsat-5 and Landsat-7 Collection 1 Level 1 images for specific studies and more recently, with the end of Landsat Collection 1 [18], BDC is obtaining Collection 2 data, for both Landsat-8 and Landsat-9.

All of the mentioned downloads and processings have been performed by continuously developing a software capable of handling optical remote sensing chains, the BDC *Collection Builder*. One of the main challenges on building this software consists in maintaining these chain process updated. Data and processors are built for specific uses. For instance, atmospheric correction processors or cloud masking algorithms heavily rely on sensor's image characteristics, for example requiring specific spectral band to exist. Due to that a processor may not be applicable to all inputs. Besides, the constant advances and improvements on these processors may require adaptations to maintain a previous established workflow. Based on that, we intend to allow multiple versions of each processor to be used by the BDC *Collection Builder*, e.g. to correct Sentinel-2 images from early 2016 using Sen2cor, only the version 2.5.5 can be used due to incompatibility with further versions (2.8.0, 2.9.0 and 2.10.1).

One of the main challenges faced by BDC project is derived from the use of BDC *Collection Builder* and BDC *Data Cube Builder* [1], which is to deal with Big Data. If we consider that 392 cells of the World Reference System Grid 2 (WRS2) used by Landsat data intersects with Brazil, there are a total of 10 TB of Landsat-8 data per year, considering only Level 1 data. When we consider the Military Grid reference System (MGRS) used by Sentinel-2, Brazil is covered by 1045 cells, which results in 60 TB per year of Level-1C compressed data. Summing this with CBERS-4 MUX ($\sim$ 4 TB/Year), CBERS-4 WFI ($\sim$ 7 TB/Year), MODIS products, CBERS-4A (WFI) and other satellite data. When considering that BDC handles more products than the mentioned Level-1, e.g. Surface Reflectance Products, data cubes, samples and classifications for several years, it sums to a few Petabytes [1].

The use of BDC *Collection Builder* allowed the BDC project to obtain more than 225 TB of Sentinel-2 L1C data and 55 TB of Landsat-8 data, as well as process them to surface reflectance products, more than doubling the occupied volume. Other sensors data, as CBERS-4, Terra and Aqua products sum even more volume to this value. The amount of acquired and processed data along the years of the BDC project, shows that BDC *Collection Builder* is a consistent

tool to handle processing chains of optical remote sensing.

For the future, it is intended to expand the types of input data and processors accepted by the software for instance acquire and process data from the remaining sensors on-board CBERS-4A and add new satellite/sensor to the workflow, as: Amazonia-1, Sentinel-2C and Sentinel-2D. Another upgrade is to include processors to perform different corrections as Nadir BRDF Adjuted Reflectance (NBAR) [19] and implement workflow for Synthetic Aperture Radar (SAR) data, as Sentinel-1. Finally, we would like to achieve a level of automation to generate on-demand ARD, to serve as inputs to on-demand data cubes.

4. REFERENCES

[1] K. R. Ferreira and et al., "Earth Observation Data Cubes for Brazil: Requirements, Methodology and Products," *Remote Sensing*, vol. 12, no. 24, pp. 4033, dec 2020.

[2] C. Kuenzer and et al., *Remote Sensing Time Series*, vol. 22 of *Remote Sensing and Digital Image Processing*, Springer International Publishing, Cham, 2015.

[3] A. Siqueira and et al., "CEOS Analysis Ready Data For Land – An Overview on the Current and Future Work," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*. jul 2019, pp. 5536–5537, IEEE.

[4] STAC, "Spatiotemporal asset catalogs," https://stacspec.org/, 2022, Access in 12 Jan. 2022.

[5] M. Silva and A. Andrade, "Geração de imagens de reflectância de um ponto de vista geométrico," *Revista Brasileira de Geomática*, vol. 1, no. 1, pp. 23, 2013.

[6] USGS, "Landsat 9," https://landsat.gsfc.nasa.gov/satellites/landsat-9/, 2022, Access in 12 Jan. 2022.

[7] USGS, "Landsat collection 2," Tech. Rep., US Geological Survey, jan 2021.

[8] USGS, "Product Guide: Landsat 8 Surface Reflectance Code (LaSRC) Product," Tech. Rep., US Geological Survey, mar 2017.

[9] USGS, "Espa and product-related code repository location changes," https://www.usgs.gov/landsat-missions/espa-and-product-related-code-repository-location-changes, 2022, Access in 12 Jan. 2022.

[10] E. Vermote and et al., "Preliminary analysis of the performance of the landsat 8/oli land surface reflectance product," *Remote Sensing of Environment*, vol. 185, pp. 46–56, 2016.

[11] M. Main-Knorn and et al., "Sen2Cor for Sentinel-2," *Image and Signal Processing for Remote Sensing*, p. 3, 2017.

[12] B. Pflug and et al., "Next updates of atmospheric correction processor Sen2Cor," *Image and Signal Processing for Remote Sensing*, vol. 1153304, no. September, pp. 2, 2020.

[13] K. Didan, "MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006 [Data set]," https://doi.org/10.5067/MODIS/MOD13Q1.006, 2022, Access in 17 Jan. 2022.

[14] S. Qiu, Z. Zhu, and B. He, "Fmask 4.0: Improved cloud and cloud shadow detection in landsats 4–8 and sentinel-2 imagery," *Remote Sensing of Environment*, vol. 231, pp. 111205, 2019.

[15] A. Sanchez and et al., "Comparison of Cloud Cover Detection Algorithms on Sentinel–2 Images of the Amazon Tropical Forest," *Remote Sensing*, vol. 12, no. 8, pp. 1284, apr 2020.

[16] VMWare, "Rabbitmq," https://www.rabbitmq.com/, 2022, Access in 13 Jan. 2022.

[17] Ask Solem, "Celery," https://docs.celeryproject.org/en/stable/index.html, 2022, Access in 13 Jan. 2022.

[18] M. A. Wulder and et al., "Current status of Landsat program, science, and applications," *Remote Sensing of Environment*, vol. 225, no. March, pp. 127–147, 2019.

[19] M. Claverie and et al., "The Harmonized Landsat and Sentinel-2 surface reflectance data set," *Remote Sensing of Environment*, vol. 219, no. September, pp. 145–161, 2018.