BUILDING EARTH OBSERVATION DATA CUBES ON AWS

Karine R. Ferreira[1]*, Gilberto R. Queiroz[1], Rennan F. B. Marujo[1], Raphael W. Costa[1]

[1] National Institute for Space Research (INPE)
Avenida dos Astronaustas, 1758, São José dos Campos, SP, Brazil.
(karine.ferreira, gilberto.queiroz, rennan.marujo, raphael.costa)@inpe.br

KEY WORDS: remote sensing images, big data, Earth observation data cubes, cloud computing, image time series analysis

ABSTRACT:

Image time series analysis and machine learning methods have been widely used in recent years to extract information from big data of remote sensing images. To support image time series analysis, remote sensing images have been modeled as Earth observation (EO) data cubes. EO data cubes can be defined as a set of time series associated to spatially aligned pixels ready for analysis. This paper describes an application for building EO data cubes on the Amazon Web Service (AWS) cloud computing environment. The *Data Cube Builder on AWS* application is based on a serverless approach to produce EO data cubes from remote sensing images stored in AWS buckets. In this work, we present the architecture of this application and its use to produce EO data cubes for Brazil from big data of remote sensing images.

1. INTRODUCTION

Cloud computing is a paradigm that provides computing as services, by solving crucial big data handling issues. Recently, cloud computing environments have emerged as suitable solutions for big Earth observation data, such as Amazon Web Services (AWS) and Google Earth Cloud (Yang et al., 2017). In recent years, Earth observation satellites have generated big volumes of images with different spatial and temporal resolutions. In 2019, the volume of open data produced by Landsat-7 and -8, MODIS (Terra and Aqua units), and Sentinel-1, -2 and -3 satellites were around 5 petabytes (Soille et al., 2018).

The AWS Public Dataset Program is an initiative that covers the costs of storage for publicly available high-value cloud-optimized data sets, including remote sensing imagery. Under this program, the Earth on AWS[1] initiative provides collections of images stored in buckets, that represent a space of storage, uniquely identified by names. Three examples of AWS resources with remote sensing images are USGS Landsat[2], Sentinel-2 [3], and CBERS-4 [4].

This paper describes an application for building Earth observation (EO) data cube on AWS, called *Data Cube Builder on AWS*. This application is open source and is under development by the Brazil Data Cube project team to produce EO data cubes from open data sets of satellite imagery under the Earth on AWS program.

2. EARTH OBSERVATION DATA CUBES FOR BRAZIL

Image time series analysis and machine learning methods have been widely used in recent years to produce land use and cover mappings from big data of remote sensing images (Gomez et al., 2016) (Picoli et al., 2020) (Bullock et al., 2020) (Simoes et al., 2021). Mainly to support image time series analysis, big data of remote sensing images are modeled as Earth observation (EO) data cubes. EO data cubes can be defined as a set of time series associated to spatially aligned pixels ready for analysis (Appel and Pebesma, 2019). They can also be defined as multidimensional arrays with spatial and temporal dimensions and spectral derived properties, created from remote sensing images.

Nowadays, there are distinct initiatives to produce EO data cubes for specific countries or regions, such as *Digital Earth Asutralia* (Lewis et al., 2017), *Swiss Data Cube* (Giuliani et al., 2017), *Africa Regional Data Cube* (Killough, 2019) and *Brazil Data Cube* (Ferreira et al., 2020). The Brazil Data Cube project is producing EO data cubes of CBERS-4, Sentinel-2 and Landsat-8 satellite images for the entire Brazilian territory (Ferreira et al., 2020). This project is developing a computational platform composed of web services, software applications and iterative computing environments to discover, access, process and analyze these EO data cubes. Using artificial intelligence, machine learning, and image time series analysis, land use and land cover maps are being produced from these EO data cubes.

Based on a hierarchical tiling system, the Brazil Data Cube project is producing two types of EO data cubes: *identity* and *temporal-composed*. This hierarchical tiling system is composed of three different grids using the Albers Equal Area projection and SIRGAS 2000 datum. These three grids have distinct tile sizes, 6×4, 3×2 and 1.5×1 degrees, referred as *BDC_LG* (large), *BDC_MD* (medium) and *BDC_SM* (small). Figure 1 shows the *BDC_SM* (small) grid that has 544 tiles of 1.5×1 degree to cover the entire country.

Identity data cubes are produced using all available images in a time interval. For each tile and date, all available images are cropped, reprojected and resampled. Therefore, the time series extracted from these data cubes may not be regular or equidistant in time, as illustrated in Figure 2. This figure shows two time series extracted from an identity data cube associated to two distinct locations, red and blue. The red time series is

* Corresponding author

[1] https://aws.amazon.com/earth

[2] https://registry.opendata.aws/usgs-landsat/

[3] https://registry.opendata.aws/sentinel-2/

[4] https://registry.opendata.aws/cbers/

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B3-2022-597-2022 | © Author(s) 2022. CC BY 4.0 License.

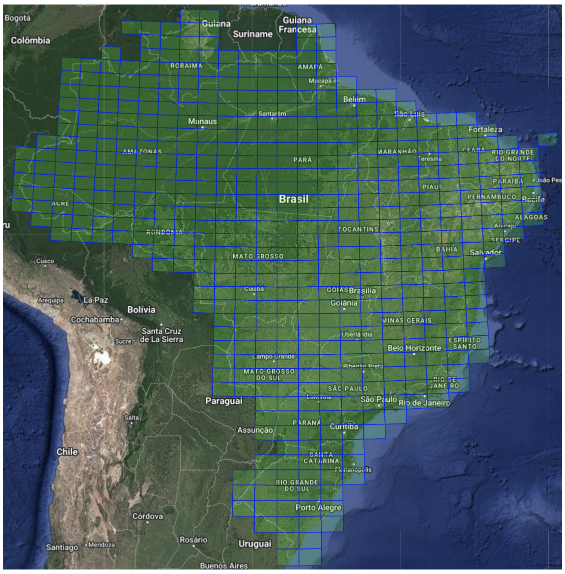


Figure 1. Tiling system of the Brazil Data Cube project.

regular in time with one observation every 5 days, but the blue time series is not.

Considering that many image time series algorithms are not prepared to handle non equidistant time series, BDC also produced *temporal-composed data cubes*. These products are regular in time and are created by using a temporal compositing function to select the best pixels (free of cloud and cloud shadow) obtained in each period (e.g. a month or 16 days). In short, the temporal-composed data cubes are products that reduce the original time dimension of the data in regular periods, for instance 16 days or monthly, trying to eliminate pixels contaminated with clouds, cloud shadows and snow. These products provides equidistant time series.

3. DATA CUBE BUILDER ON AWS

Brazil is a huge country, with a territory of over 8.5 millions of km^2. To build EO data cubes for the entire country from 2015 to 2023, the Brazil Data Cube project has to process around 734 Terabytes (TB) of Sentinel-2 surface reflectance (SR) images, 108 TB of Landsat-8 SR images and 108 TB of CBERS-4 (MUX and WFI) SR images, as shown in Figure 3. Sentinel-2 SR is the biggest data set due to its spatial resolution, temporal revisit and number of spectral bands. The volume of EO temporal-composed data cubes produced and handled by the Brazil Data Cube project is presented in Figure 4. We are producing 16-days temporal-composed data cubes of Sentinel-2, Landsat-8 and CBERS-4 WFI images and monthly temporal-composed data cubes of CBERS-4 MUX images.

Cloud computing environments are suitable for processing these big data sets efficiently. These environments do not impose any API for data processing, allowing the development of open source technologies in non-proprietary scripting languages. In the Brazil Data cube project, we are using the AWS to produce part of the 16-days temporal-composed data cubes of Sentinel-2 images for Brazil. The AWS was chosen due to two reasons: (1) Sentinel-2 SR data sets are already stored and available on AWS under the Open Data program [5] [6]; and (2) AWS provides Cloud Credits for Research Program [7] for doing research using Earth Observation or other geospatial data on AWS.

The Data Cube Builder on AWS application is based on a serverless approach, using the AWS Lambda service. Its architecture is shown in Figures 5. A serverless ecosystem allows to define workloads to achieve scalability, performance and cost efficiency, without managing the underlaying infrastructure. These workloads scale thousands of concurrent processing per second, which enables to generate EO data cube on demand for any area in short time.

The AWS Lambda service defines *Functions* to be executed along the workload. These *Functions* have quotas, which limit the amount of compute and storage resources in executions. These lambda functions are dispatched by *Simple Queue Service (SQS)* and store the metadata result into *DynamoDB*, a fast and flexible NoSQL database service to deal with concurrent requests. At the end of processing, the data cubes produced are stored in *Simple Storage (S3)* and their definitions into *Relational Database Service (RDS)*.

Some interfaces of the Data Cube Builder on AWS are shown in Figures 6, 7 and 8. Figure 6 illustrates the definition of a new data cube, which includes specifying a source Spatial Temporal Asset Catalog (STAC) that will be used to consult and obtain input images to generate the data cube, the image collection in this STAC, the satellites sensors that will be used, a date range of the period to be considered, the AWS bucket name to store the produced data cubes as well as information about the temporal composition and tiling system or grid. Users can also select or define vegetation indices to be produced such as Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI).

Before creating the data cubes, its metadata can be verified through the Preview interface, as illustrated in Figure 7. The STAC is used to consult all available images for each cell grid, also referred as tiles, according to the temporal composition and temporal range defined. The result of this process are temporal-composed data cubes. The main advantage on executing the processing using AWS consists in scaling the generation of the EO data cubes, allowing the independent processing of each item, which consists in a time step of a tile, allowing the parallel and distributed processing of all tiles and dates simultaneously, restricted to AWS configuration.

The Cube Builder on AWS also contains a graphical interface to verify and validate the produced data cubes, as illustrated in Figure 8. Following this interface, each date, in each processed tile, can be visually inspected. Besides that, a user can even reprocess specific items, e.g. when input images have been through a new correction processing, reprocess or update, or when a defective data has been used as input and should be removed.

We have two versions of the Data Cube Builder application, one that runs on AWS, described in this paper, and another that runs in local environments. Both versions are constantly evolving and are synchronized, implementing new features as new requirements for the Brazil Data Cube products appear. This includes product validation, software optimization and several others temporal compositing functions beyond the best pixel approach, as Median and Average (Ferreira et al., 2020) and others for specific applications, for instance time-priority composition, used to build visual mosaics.

[5] https://registry.opendata.aws/sentinel-2/

[6] https://registry.opendata.aws/sentinel-2-l2a-cogs/

[7] https://aws.amazon.com/earth/research-credits/

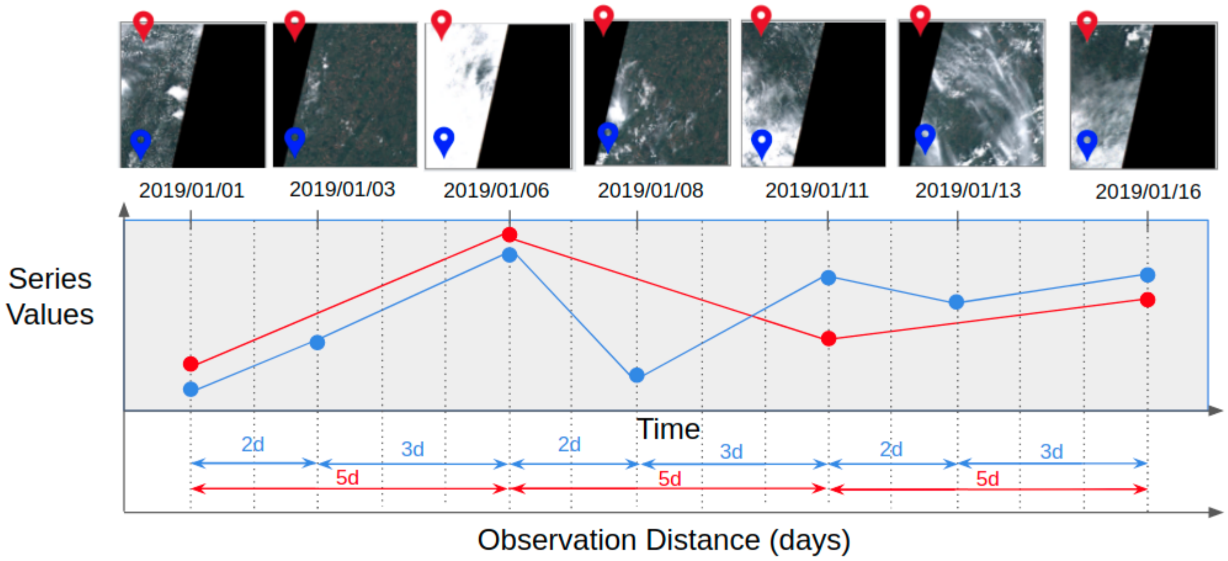


Figure 2. Time series extracted from identity EO data cubes can be regular in time (red time series) or not (blue time series).

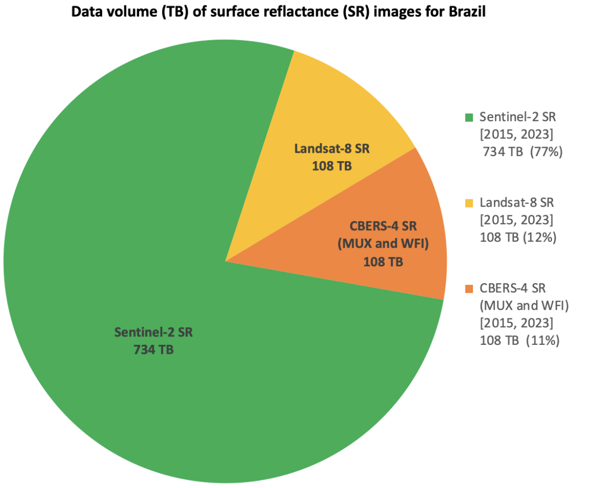


Figure 3. Data volume (TB) of surface reflectance images for Brazil

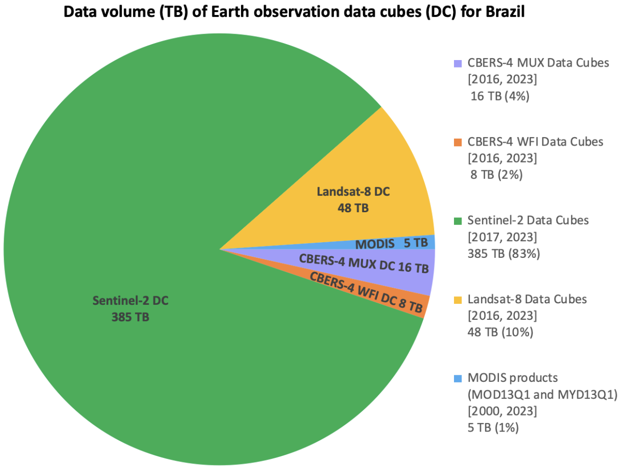


Figure 4. Data volume (TB) of Earth observation data cubes for Brazil

4. RESULTS AND FINAL REMARKS

Table 1 presents the time elapsed and cost to produce EO data cubes using the Data Cube Builder on AWS application. We used this application to produce two years (2017 and 2018) of 16-days temporal-composed EO data cubes of Sentinel-2 (A and B) images for the entire Brazilian territory, based on the grid with 544 tiles of 1.5 by 1 degree each, shown in Figure 5. The Data Cube Builder on AWS application took 13 days to produce these data cubes and the costs, shown in Table 1, were paid by the credits earned under the GEO (Group on Earth Observations) and AWS Earth Observation Cloud Credits Program. These temporal-composed EO data cubes of Sentinel-2 images are regular in time, containing the best pixels obtained for each 16 days, and are available as Open Data on AWS [8]. The Data Cube Builder on AWS application takes 30 minutes to process one year of Sentinel-2 images for each single tile, as shown in Table 1.

Tiles	Period	Time elapsed	Cost (U$)
1	1 year	30 mins	7.00
90	1 year	18 hours	630.00
544	2 year	13 days	7616.00

Table 1. Time elapsed and cost to produce EO data cubes using the Data Cube Builder on AWS.

The Data Cube Builder on AWS is a serverless application optimized for cloud computing environments. This is its main advantage over other solutions for EO data cube production, such as *gdalcubes* (Appel and Pebesma, 2019) and *Data Cube on Demand (DCoD)* (Giuliani et al., 2020). Appel and Pebesma (2019) present a use case study using *gdalcubes* on a local machine and remark that the rational trend for large data sets is to move computations to cloud platforms where the data is already available. They argue that its is possible to run several of *gdalcubes* worker instances in containerized cloud environments to allow process distribution over many compute instances.

[8] https://registry.opendata.aws/brazil-data-cubes/

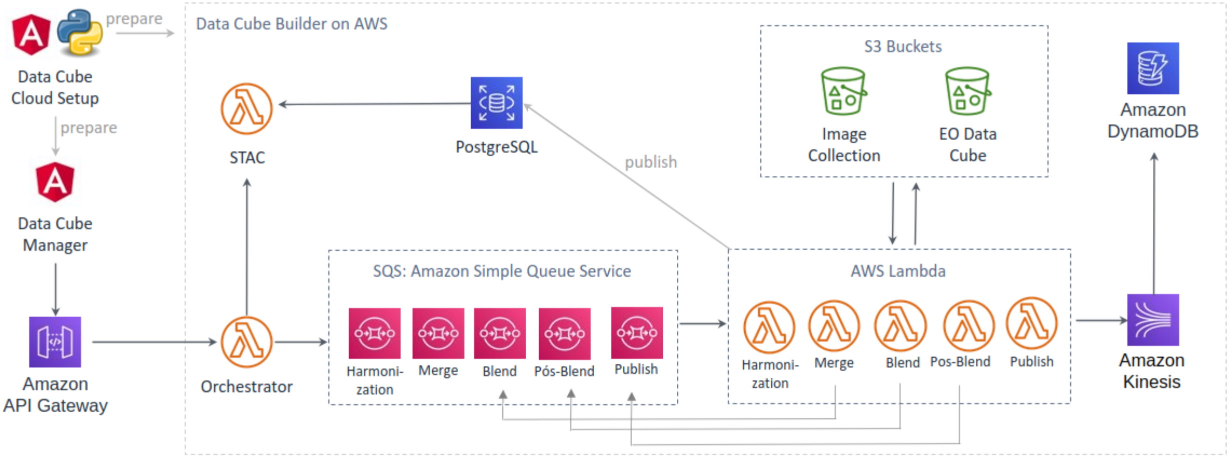
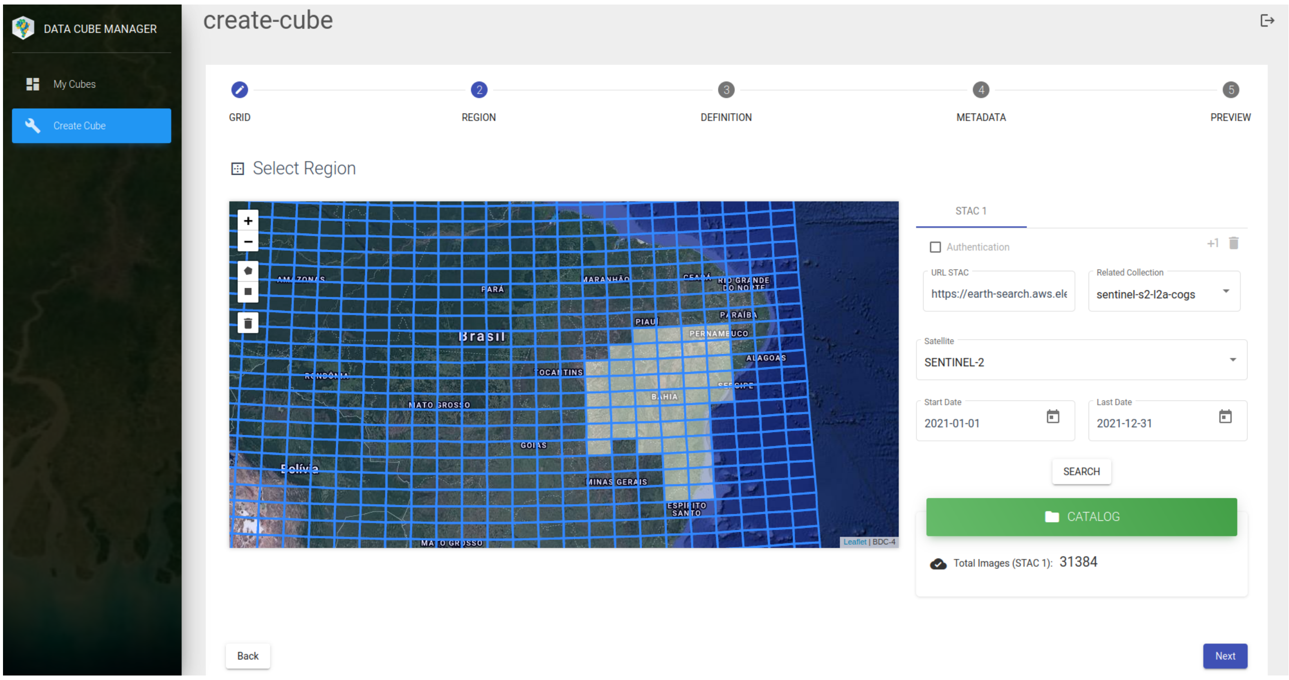


Figure 5. Architecture of Data Cube Builder on AWS.



☐ Authentication

URL STAC: https://earth-search.aws.ele

Related Collection: sentinel-s2-l2a-cogs

Satellite: SENTINEL-2

Start Date: 2021-01-01

Last Date: 2021-12-31

Total Images (STAC 1): 31384

Figure 6. Graphical interface of Data Cube Builder on AWS - Data Cube Definition.

The *DCoD* is a solution to automate the generation of an Open Data Cube (ODC) instance virtually in distinct infrastructures, including the process of selecting, ordering, downloading and ingesting Analysis Ready Data (ARD) of remote sensing images (Giuliani et al., 2020). Giuliani et al. (2020) argue that there are ODC instances already running on AWS and so it should be possible to deploy the DCoD approach on this type of environment too.

5. ACKNOWLEDGEMENTS

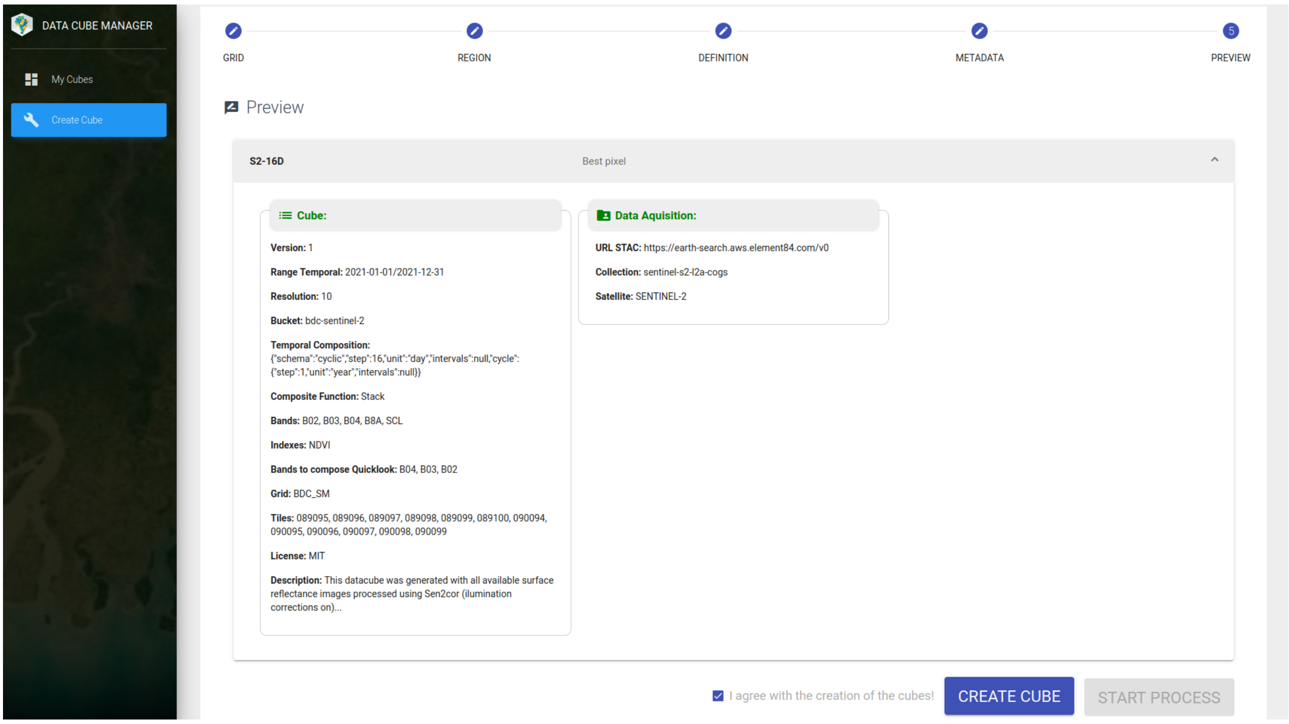
This study was financed in part by the Coordenação de Aperfeicoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001; by the Amazon Fund through the financial collaboration of the Brazilian Development Bank (BNDES) and the Foundation for Science, Technology and Space Applications (FUNCATE) to develop the project *Environmental Monitoring of Brazilian Biomes* and its subproject *Brazil Data Cube* (process number 17.2.0536.1); and by the GEO (Group on Earth Observations) and AWS Earth Observation Cloud Credits Program to develop the project *Brazil Data Cube using AWS for Land Use and Cover Change*.

REFERENCES

Appel, M., Pebesma, E., 2019. On-demand processing of data cubes from satellite image collections with the gdalcubes library. *Data*, 4(3), 92.

Bullock, E. L., Woodcock, C. E., Olofsson, P., 2020. Monitoring tropical forest degradation using spectral unmixing and Landsat time series analysis. *Remote sensing of Environment*, 238, 110968.

Ferreira, K. R., Queiroz, G. R., Vinhas, L., Marujo, R. F. B., Simoes, R. E. O., et. al, 2020. Earth Observation Data Cubes



Preview

S2-16D   Best pixel

Cube:

Version: 1

Range Temporal: 2021-01-01/2021-12-31

Resolution: 10

Bucket: bdc-sentinel-2

Temporal Composition: {"schema":"cyclic","step":16,"unit":"day","intervals":null,"cycle":{"step":1,"unit":"year","intervals":null}}

Composite Function: Stack

Bands: B02, B03, B04, B8A, SCL

Indexes: NDVI

Bands to compose Quicklook: B04, B03, B02

Grid: BDC_SM

Tiles: 089095, 089096, 089097, 089098, 089099, 089100, 090094, 090095, 090096, 090097, 090098, 090099

License: MIT

Description: This datacube was generated with all available surface reflectance images processed using Sen2cor (ilumination corrections on)...

Data Aquisition:

URL STAC: https://earth-search.aws.element84.com/v0

Collection: sentinel-s2-l2a-cogs

Satellite: SENTINEL-2

☑ I agree with the creation of the cubes!

CREATE CUBE   START PROCESS

Figure 7. Graphical interface of Data Cube Builder on AWS - Data Cube Preview.

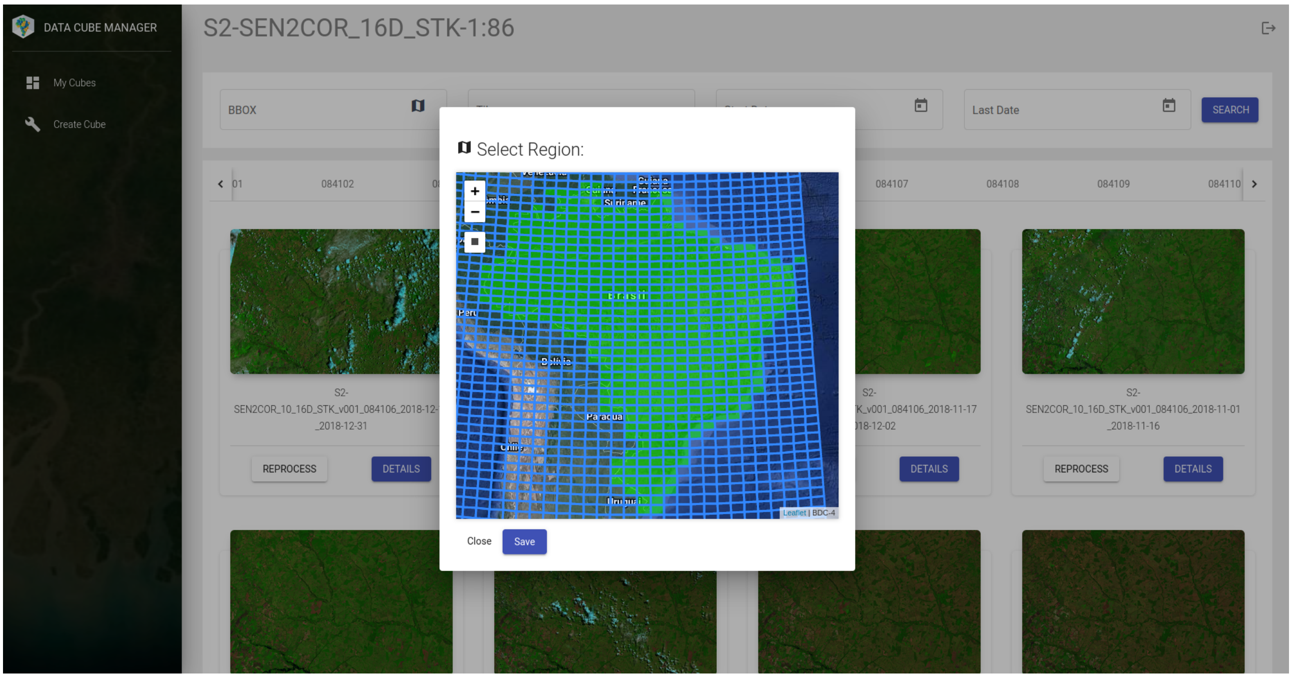


Figure 8. Graphical interface of Data Cube Builder on AWS - Data Cube Verification.

for Brazil: Requirements, Methodology and Products. *Remote Sensing*, 12(24), 4033.

Giuliani, G., Chatenoux, B., De Bono, A., Rodila, D., Richard, J.-P., Allenbach, K., Dao, H., Peduzzi, P., 2017. Building an Earth Observations Data Cube: Lessons Learned from the Swiss Data Cube (SDC) on Generating Analysis Ready Data (ARD). *Big Earth Data*, 1(1-2), 100–117.

Giuliani, G., Chatenoux, B., Piller, T., et al., 2020. Data Cube on Demand (DCoD): Generating an earth observation Data Cube anywhere in the world. *International Journal of Applied Earth Observation and Geoinformation*, 87, 102035.

Gomez, C., White, J. C., Wulder, M. A., 2016. Optical Remotely Sensed Time Series Data for Land Cover Classification: a Review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 116, 55–72.

Killough, B., 2019. The Impact of Analysis Ready Data in the Africa Regional Data Cube. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, Yokohama, Japan, 5646–5649.

Lewis, A., Oliver, S., Lymburner, L., Evans, B., Wyborn, L., Mueller, N., Raevksi, G., Hooke, J., Woodcock, R., Sixsmith, J., Wu, W., Tan, P., Li, F., Killough, B., Minchin, S., Roberts, D., Ayers, D., Bala, B., Dwyer, J., Dekker, A., Dhu, T., Hicks, A., Ip, A., Purss, M., Richards, C., Sagar, S., Trenham, C., Wang, P., Wang, L.-W., 2017. The Australian Geoscience Data Cube — Foundations and Lessons Learned. *Remote Sensing of Environment*, 202, 276–292.

Picoli, M. C. A., Simoes, R., Chaves, M., Santos, L. A., Sanchez, A., Soares, A., Sanches, I. D., Ferreira, K. R., Queiroz, G. R., 2020. CBERS DATA CUBE: A POWERFUL TECHNOLOGY FOR MAPPING AND MONITORING BRAZILIAN BIOMES. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-3-2020, 533–539. https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/V-3-2020/533/2020/.

Simoes, R., Camara, G., Queiroz, G., Souza, F., Andrade, P. R., Santos, L., Carvalho, A., Ferreira, K., 2021. Satellite Image Time Series Analysis for Big Earth Observation Data. *Remote Sensing*, 13(13), 2428.

Soille, P., Burger, A., De Marchi, D., Kempeneers, P., Rodriguez, D., Syrris, V., Vasilev, V., 2018. A versatile data-intensive computing platform for information retrieval from big geospatial data. *Future Generation Computer Systems*, 81, 30–40. https://doi.org/10.1016/j.future.2017.11.007.

Yang, C., Yu, M., Hu, F., Jiang, Y., Li, Y., 2017. Utilizing cloud computing to address big geospatial data challenges. *Computers, environment and urban systems*, 61, 120–128.