Integração dos ambientes Brazil Data Cube e Open Data Cube

Felipe Menino Carlos[1], **Vitor C. F. Gomes**[2], **Gilberto Ribeiro de Queiroz**[1], **Karine Reis Ferreira**[1], **Rafael Santos** [1]

[1]Instituto Nacional de Pesquisas Espaciais (INPE)
12227-010 – São José dos Campos – SP – Brasil

[2] Divisão de C4ISR – IEAv/DCTA
12.228-001 – São José dos Campos – SP – Brasil

{felipe.carlos,gilberto.queiroz,karine.ferreira}@inpe.br
rafael.santos@inpe.br, vitorvcfg@fab.mil.br

Abstract. *The availability of large quantities of Earth observation data by satellite images has enabled different technologies and research. Given the challenges that this volume of data presents, the organization of these in data cube formats becomes fundamental for any large scale study. Providing tools that use these data cubes simpler allows their widespread use in various contexts. This article aims to present the ongoing work on the extension of the data cube analysis and visualization tools produced by the Brazil Data Cube project by integrating these tools with the Open Data Cube framework.*

Resumo. *A disponibilização de grandes quantidades de dados de observação da Terra por imagens de satélite tem possibilitado o desenvolvimento de diferentes tecnologias e pesquisas. Frente aos desafios que esse volume de dados apresenta, a organização desses em formatos de cubos de dados passa a ser fundamental. Disponibilizar ferramentas que tornem o uso desses cubos de dados mais simples, permite sua ampla utilização em diversos contextos. O presente artigo tem por objetivo apresentar o trabalho em andamento da extensão das ferramentas de analise e visualização dos cubos de dados produzidos pelo projeto Brazil Data Cube através da integração dessas com o framework Open Data Cube.*

1. Introdução

Nos últimos anos, a ciência e a indústria geoespacial têm desenvolvido inúmeras aplicações inovadoras graças à maior disponibilidade de dados de observação da Terra (EO, do inglês *Earth Observation*). Tanto os avanços tecnológicos dos equipamentos de coleta de dados quanto de armazenamento, aliadas à adoção de políticas de dados abertos pelas agências espaciais, têm propiciado a criação dessas aplicações [Soille et al. 2018]. No entanto, lidar com esses conjuntos massivos de dados ainda representa um grande desafio para a extração de todo o potencial e valor destes [Appel and Pebesma 2019]. Frequentemente, esses dados excedem as capacidades de memória, armazenamento e processamento de computadores tradicionalmente utilizados para esta finalidade [Câmara et al. 2014].

Para lidar com estes desafios, a comunidade científica tem utilizado o conceito de *Earth Observation Data Cube* (EODC), que através de conjuntos especializados de tecnologias, busca solucionar os problemas com grandes volumes de dados [Giuliani et al. 2020]. Uma das principais plataformas que tem se destacado nesse cenário é o Open Data Cube (ODC) [Gomes et al. 2020], que possui várias ferramentas e serviços para a análise e gerenciamento de grandes volumes de dados e que vem sendo utilizada por diversas iniciativas e instituições ao redor do mundo [Dhu et al. 2019].

No contexto brasileiro, o projeto Brazil Data Cube (BDC) é uma iniciativa criada em 2019, pelo Instituto Nacional de Pesquisas Espaciais (INPE), que tem por objetivo produzir cubos de dados multidimensionais para todo o território brasileiro, possibilitando a geração e análise de informações sobre o uso e cobertura do solo através de diferentes métodos, como a análise de séries temporais e uso de algoritmos de Aprendizado de Máquina. Para realizar suas atividades, o projeto BDC atualmente utiliza tecnologias de *Big Data* e ambientes de computação em nuvem para processar os cubos de dados [Ferreira et al. 2020].

Esse artigo apresenta o trabalho em andamento de integração entre os produtos de *software* do projeto BDC e a plataforma Open Data Cube. O objetivo dessa integração é possibilitar o acesso, processamento e análise das coleções de imagens e dos cubo de dados gerados pelo projeto BDC na plataforma ODC. Essa integração amplia os serviços e ferramentas que podem ser utilizadas para acessar, visualizar e analisar os dados produzidos pelo projeto BDC.

2. Open Data Cube (ODC)

O ODC é um *framework* analítico composto por uma série de estruturas de dados e ferramentas que facilitam o gerenciamento e análise de dados de EO [Gomes et al. 2020]. Ele permite a catalogação de conjuntos massivos de dados *raster*, possibilitando também o trabalho com dados que possuem alta dimensionalidade temporal. O ecossistema do ODC é composto pelos seguintes componentes [Open Data Cube 2019]:

- **Ferramentas de linha de comando**: Ferramentas para o gerenciamento dos dados registrados no ODC;
- **ODC Explorer**: Interface *web* que possibilita aos usuários explorar e buscar os dados que estão registrados no ODC;
- **ODC Stats**: Aplicação que facilita a análise estatística de grandes conjuntos de dados *raster*;
- **Web User Interface**: Interface para a visualização interativa dos resultados de execução dos algoritmos de análise;
- **OGC Web Services**: Serviços que possibilitam o uso interoperável dos dados registrados no ODC; e
- **Interface de programação (API)**: API em linguagem Python que possibilita a busca, acesso, análise e visualização dos dados, podendo ser utilizado junto ao ambiente interativo `Jupyter Notebook`.

Além desses, o ecossistema do ODC possui o componente `ODC Core`, uma camada entre os conjuntos de dados e os componentes citados anteriormente, responsável por fornecer uma estrutura analítica e de catalogação, capaz de lidar com um conjunto massivo de imagens de Sensoriamento Remoto. Para seu funcionamento, o ODC Core

utiliza o conceito de indexação, que é o processo responsável por catalogar os dados que estarão disponíveis para uso. Nesse processo, são registrados os metadados das imagens e os locais de armazenamento, que pode ser um sistema de arquivos distribuídos ou um serviço de armazenamento na nuvem. Uma vez registrados, os dados podem ser consumidos pelos demais componentes do ecossistema ODC.

3. Integração BDC-ODC

Com o objetivo de permitir que os produtos de dados gerados pelo BDC sejam disponibilizados por meio das ferramentas e serviços do ODC, faz-se necessária a integração dos dois ambientes. Para o processo de integração, três fases foram definidas. A primeira delas trata da indexação dos dados do BDC dentro do ODC Core. A segunda fase realiza a seleção, integração e configuração dos serviços e ferramentas do ODC, de modo que essas permitam o uso dos dados do BDC considerando suas características específicas. Por fim, na terceira fase, faz-se a criação de um projeto de infraestrutura computacional que permita aos usuários consumir facilmente as ferramentas e facilidades resultantes dessa integração. Os passos realizados em cada uma dessas etapas são especificados nas próximas subseções.

3.1. Indexação

O primeiro passo necessário para a integração entre os ambientes é a indexação dos produtos de dados disponibilizados pelo BDC no catálogo do ODC. Para isso, inicialmente foi feita a configuração de um *container* Docker com o ODC Core. Em seguida, foi desenvolvida uma ferramenta de indexação, nomeada de `stac2odc` [1], responsável por fazer a busca e recuperação dos dados disponíveis no catálogo BDC-STAC [Zaglia et al. 2019] e registrá-los no catálogo de imagens do ODC Core. A Figura 1 ilustra esse processo.

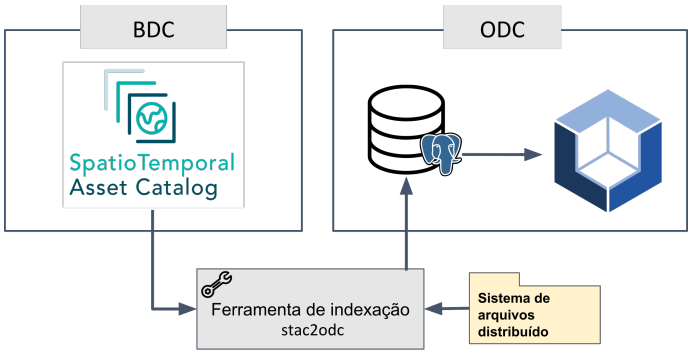


Figura 1. Fluxo de indexação dos dados BDC-ODC

Para evitar a movimentação e replicação de dados, o ODC Core tem acesso direto ao sistema de arquivos distribuídos utilizado no projeto BDC, o qual é utilizado pela ferramenta `stac2odc` durante a indexação. Para usos em que não esteja disponível acesso direto à infraestrutura, a ferramenta `stac2odc` disponibiliza a opção de *download* dos dados durante a indexação.

[1]Disponível em: `https://github.com/brazil-data-cube/bdc-odc`

3.2. Ferramentas e serviços integrados

Com o processo de indexação realizado, por padrão, os dados do BDC podem ser gerenciados através das ferramentas de linha de comando ou serem acessados pela API em Python, ambas disponíveis no ecossistema ODC.

Para permitir a visualização e recuperação facilitada aos dados indexados no ODC, optou-se, nesse trabalho, pela implantação das ferramentas `ODC-Explorer` e `OGC Web Services`. Diferente da primeira fase, o processo de implantação dessas ferramentas exigiu configurações específicas para o funcionamento com os dados do BDC. Nessa etapa, foi necessária a modificação do código fonte dessas ferramentas, pois elas não estavam preparadas para lidar com dados gerados em grades espaciais de referência customizadas, que é o caso dos dados produzidos no BDC. Essas alterações estão disponíveis nos repositórios de código `datacube-explorer` e `datacube-ows` do BDC[2].

3.3. Infraestrutura

Para facilitar o uso das ferramentas analíticas disponibilizadas com integração do ODC e BDC, fez-se a definição de um projeto de infraestrutura computacional, ilustrada pela Figura 2. A ferramenta JupyterHub disponibiliza ambientes interativos, para que cada usuário, através de um navegador *web*, possa acessar e utilizar o ambiente, sem a necessidade de realizar nenhum tipo de configuração de *software* ou movimentação de dados.

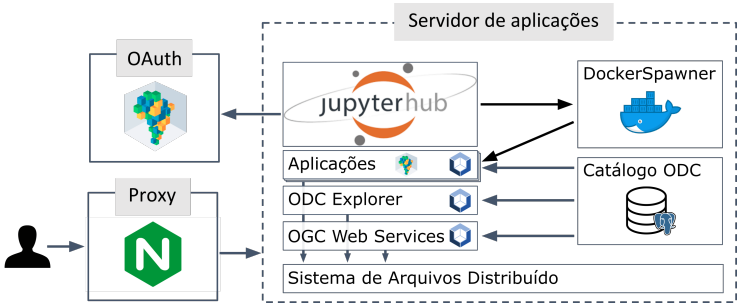


Figura 2. Infraestrutura computacional BDC-ODC

Na arquitetura implementada, o usuário pode realizar acessos ao ambientes através da autenticação no serviço *OAuth* criado no BDC. Uma vez autenticado, o usuário tem à sua disposição um ambiente de Jupyter Notebooks, criado com o uso de *containers* Docker. Por fim, estando dentro deste ambiente, o usuário pode utilizar as ferramentas de análise do ODC e BDC para realizar o processamento e análise dos dados.

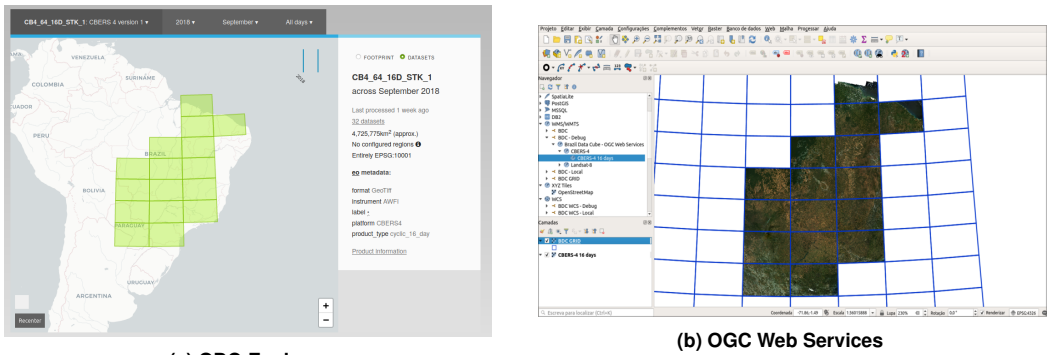
4. Resultados

Nesta seção, são apresentados exemplos de uso dos serviços e ferramentas que foram implementados nesse trabalho de integração.

A customização e configuração dos *OGC Web Services* tornou possível o consumo dos produtos de dados do BDC através dos serviços WCS, WMS e WMTS. Além disso,

[2]Disponível em: `https://github.com/brazil-data-cube`

todos os dados que estão registrados no catálogo ODC podem ser facilmente encontrados com as buscas visuais oferecidas pelo ODC-Explorer. As Figuras 3a e 3b mostram os cubos de dados gerados a partir de imagens do satélite CBERS-4, para todo o bioma Cerrado, sendo apresentado na ferramenta ODC Explorer e consumido via OGC WMS no *software* QGIS.



(a) ODC-Explorer **(b) OGC Web Services**

Figura 3. Interface do ODC-Explorer e OGC Web Services

A Figura 4 apresenta a janela de seleção de ambientes oferecida pelo JupyterHub aos usuários, bem como exemplos[3] de Jupyter Notebooks que podem ser carregados para análise dos dados após a seleção do ambiente com as ferramentas ODC.

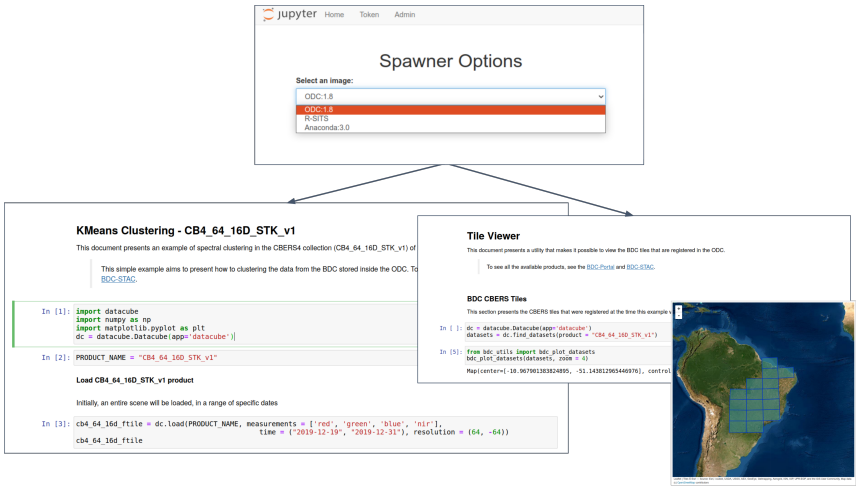


Figura 4. Interface de usuário do JupyterHub

Cabe notar que o projeto de infraestrutura computacional apresentado, permite o uso de outros ambientes além do ODC, como pode ser visto na Figura 4.

5. Considerações finais

Este artigo apresentou o trabalho em andamento da integração entre os ambientes Brazil Data Cube (BDC) e do framework Open Data Cube (ODC). A integração inicial, feita

[3]Exemplos disponíveis em: `https://github.com/brazil-data-cube/bdc-odc`

com algumas das ferramentas do ecossistema do ODC, mostraram que a realização deste processo fornece aos usuários novas formas de consumo dos produtos de dados do BDC, representando mais opções e flexibilidade aos usuários que vão consumir esses produtos.

O trabalho apresentou também o projeto de uma infraestrutura computacional que pode ser utilizada pelos usuários para que as ferramentas de análise, disponíveis após o processo de integração, sejam facilmente consumidas junto aos produtos de dados do BDC, sem a necessidade de nenhuma configuração ou movimentação de dados para os equipamentos dos usuários.

Como trabalho futuro, será feita a integração dos demais componentes do ODC no ecossistema do BDC, além da adição de facilidades na interface da plataforma, como opções para compartilhamento de resultados e publicação de dados gerados.

Agradecimento

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e ao subprojeto Brazil Data Cube do Projeto Monitoramento Ambiental dos Biomas Brasileiros, financiado com recursos do Fundo Amazônia, por meio da colaboração financeira BNDES e FUNCATE nº 17.2.0536.1

Referências

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

Câmara, G., Egenhofer, M. J., Ferreira, K., Andrade, P., Queiroz, G., Sanchez, A., Jones, J., and Vinhas, L. (2014). Fields as a Generic Data Type for Big Spatial Data. *Geographic Information Science*, page in press.

Dhu, T., Giuliani, G., Juárez, J., Kavvada, A., Killough, B., Merodio, P., Minchin, S., and Ramage, S. (2019). National Open Data Cubes and Their Contribution to Country-Level Development Policies and Practices. *Data*, 4(4):144.

Ferreira, K. R., Queiroz, G. R., Camara, G., Souza, R. C. M., Vinhas, L., Marujo, R. F. B., Simoes, R. E. O., Noronha, C. A. F., Costa, R. W., Arcanjo, J. S., Gomes, V. C. F., and Zaglia, M. C. (2020). Using remote sensing images and cloud services on aws to improve land use and cover monitoring. In *2020 IEEE Latin American GRSS ISPRS Remote Sensing Conference (LAGIRS)*, pages 558–562.

Giuliani, G., Chatenoux, B., Piller, T., Moser, F., and Lacroix, P. (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.

Gomes, V. C. F., Queiroz, G. R., and Ferreira, K. R. (2020). An Overview of Platforms for Big Earth Observation Data Management and Analysis. *Remote Sensing*, 12(8):1253.

Open Data Cube (2019). ODC: Architecture and Ecosystem - A High-Level Overview. Technical report, Open Data Cube.

Soille, P., Burger, A., De Marchi, D., Kempeneers, P., Rodriguez, D., Syrris, V., and Vasilev, V. (2018). A versatile data-intensive computing platform for information retrieval from big geospatial data. *Future Generation Computer Systems*, 81:30–40.

Zaglia, M. C., Vinhas, L., Queiroz, G. R., and Simoes, R. E. O. (2019). Catalogação de metadados do cubo de dados do Brasil com o SpatioTemporal asset catalog. *Simpósio Brasileiro De Geoinformática (GEOINFO)*, pages 280–285.