Análise de ferramentas para processamento de grandes volumes de dados espaço-temporais

Fabiana Zioti[1], Giberto Ribeiro de Queiroz[1], Karine Reis Ferreira[1]

[1]Instituto Nacional de Pesquisas Espaciais (INPE)
Caixa Postal 515 – Av. dos Astronautas, 1758, Jardim da Granja – SP – Brasil

fabi.zioti@gmail.com, {gilberto.queiroz,karine.ferreira}@inpe.br

Resumo. *Dados espaciais desempenham um papel crucial em estudos socioambientais para definições de políticas e práticas públicas que diminuam o impacto das atividades humanas sobre o meio ambiente. Atualmente, o grande volume de dados espaço-temporais e de imagens de observação da Terra trazem novos desafios às diversas áreas da ciência, em especial à computação. Neste contexto, esse trabalho apresenta uma análise das ferramentas computacionais SpatialHadoop, ST-Hadoop e Geospark para processar grandes volumes de dados espaço-temporais. Essa análise foi realizada através de um experimento com dados produzidos por projetos de monitoramento ambiental do Instituto Nacional de Pesquisas Espaciais (INPE).*

1. Introdução

Diante das mudanças observadas no planeta, com recursos naturais cada vez mais escassos, é importante fomentar estudos socioambientais e definir políticas e práticas públicas para o processo de tomada de decisões que diminuam o impacto das atividades humanas sobre o meio ambiente. O mapeamento da dinâmica do uso e cobertura da Terra tem sido considerado de grande importância para entender os efeitos das atividades humanas sobre o planeta e assim obter informações úteis para diversas áreas: gestão de recursos naturais, monitoramento ambiental, mudanças climáticas, entre outras [Foley et al. 2005]. Neste cenário, os dados espaciais desempenham um papel crucial. As imagens de sensoriamento remoto, por exemplo, tornaram-se uma importante fonte de dados espaciais empregadas no monitoramento da Terra em escala regional e global [Arvor et al. 2011, Aguiar et al. 2010, Gómez et al. 2016].

Os avanços nas tecnologias de sensoriamento remoto têm possibilitado a aquisição de dados com resoluções espaciais e temporais cada vez mais finas. Com isso, existe hoje uma grande quantidade e diversidade de dados de sensoriamento remoto disponíveis para utilização em diversas áreas. Embora a disponibilidade de grandes volumes de dados espaço-temporais proporcionam avanços em pesquisas e aplicações, o armazenamento, acesso e processamento desses dados se tornam um desafio computacional [CEOS 2018]. Ferramentas atuais para processar grandes volumes de dados espaço-temporais incluem tecnologias de propósito geral como Apache Spark[1], Apache Hadoop[2], Apache Storm[3]. Além de construção de novos sistemas ou desenvolvimento de extensões para os sistemas distribuídos como SpatialHadoop, Geospark.

[1]https://spark.apache.org/
[2]https://hadoop.apache.org/
[3]https://storm.apache.org/

Esse trabalho apresenta uma análise das ferramentas computacionais SpatialHadoop, ST-Hadoop e Geospark para processar grandes volumes de dados espaço-temporais. Essa análise foi realizada através de um experimento com dados produzidos por projetos de monitoramento ambiental do Instituto Nacional de Pesquisas Espaciais (INPE), PRODES[4] (Projeto de Monitoramento do Desmatamento na Amazônia Legal por Satélite), DETER[5](Detecção de Desmatamento em Tempo Real), TerraClass[6] e Programa Queimadas[7].

2. Tecnologias para *Big Data*

Hadoop e Spark são exemplos de tecnologias para processamento de *Big Data*. Entretanto essas ferramentas não suportam de maneira nativa dados espaço-temporais. Visando resolver essa lacuna, diversas extensões foram desenvolvidas. A Figura 1 apresenta a linha do tempo das extensões propostas para processar de forma nativa dados espaciais ou espaço-temporais para as tecnologia Hadoop e Spark.

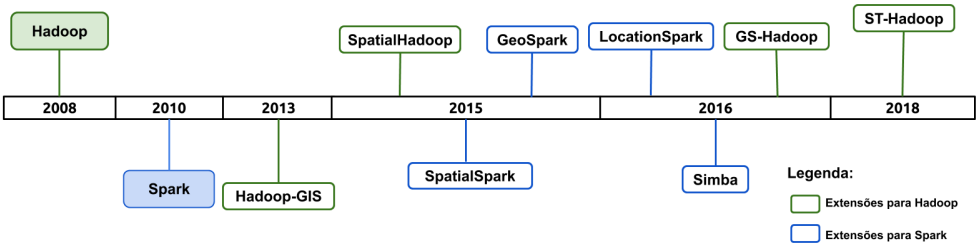

Figura 1. Linha do tempo das extensões do Hadoop e Spark.

Cada extensão possui suas próprias características e diferentes funcionalidades para lidar com grandes volumes de dados espaciais ou espaço-temporais. Pelo fato das extensões serem desenvolvidas sob diferentes estruturas, elas herdam as vantagens e desvantagens de cada uma. A Tabela 1 apresenta um comparativo de características presentes nas extensões para Hadoop e Spark.

Pandey et al. [Pandey et al. 2018] realizaram uma análise comparativa de algumas extensões baseadas no Spark. Os autores apresentam o GeoSpark como a extensão mais completa. No trabalho de [Lenka et al. 2016] é apresentada uma visão geral das arquiteturas do SpatialHadoop e GeoSpark. Eles apresentam um comparativo de tempo de execução das duas ferramentas, mas não mostra detalhes de quais operações foram comparadas. Como conclusão apresenta que o Geospark é mais rápido comparado ao SpatialHadoop, porém possui uma comunidade para suporte limitada. O trabalho de [García-García et al. 2017] apresenta uma análise das extensões SpatialHadoop e LocationSpark. Os autores avaliam a performance de dois algoritmos de *distance join queries* e apontam a extensão LocationSpark vencedora com relação ao tempo total de execução. Porém é enfatizado que o SpatialHadoop possui um tempo maior dedicado ao desenvolvimento, e se mostra mais maduro.

[4]http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes
[5]http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/deter
[6]https://www.terraclass.gov.br/
[7]http://queimadas.dgi.inpe.br/queimadas/portal

Tabela 1. Comparativo das extensões. Adaptado de [Alam et al. 2018]

Características	SpatialSpark	GeoSpark	LocationSpark	SpatialHadoop	Hadoop-GIS
Dados de Entrada	WKT	CSV, TSV, WKT, WKB, GeoJSON e Shapefile	WKT	WKT	WKT
Linguagem de Alto Nível	não possui	SQL	não possui	Pigeon	HiveQL com suporte espacial
Indexação	R-tree	R-tree, Quad-tree	R-tree, Quad-tree, IR-tree	Grid file, R-tree e R+-tree	R*-Tree , Hilbert R-Tree
Operações	Range Query, Broadcast Join e Partitioned Join	Spatial Range, Join e KNN query	Range Search, kNN, Spatio-Textual, Spatial-Join e k-NN Join	Range Query, k-NN, Spatial-Join	Range, Nearest Neighbor e Spatial-Join

A extensão ST-Hadoop é tida como a primeira ferramenta a dar suporte nativo para dados espaço-temporais. No entanto não existe na literatura uma comparação dessa e outras ferramentas especificamente para dados espaço-temporais.

3. Dados espaço-temporais produzidos pelo INPE

O projeto DETER mapeia os alertas de desmatamento em tempo real da Amazônia brasileira desde 2004 [Diniz et al. 2015]. São produzidos diariamente dados vetoriais com tempo de observação associado. O PRODES é o projeto que monitora o desmatamento por corte raso na Amazônia brasileira desde 1988 e no bioma Cerrado desde 2016, fornecendo taxas e dados vetoriais anuais referentes ao desmatamento para estas regiões [INPE 2019a].

O Projeto TerraClass desenvolvido pelo INPE em parceria com a Embrapa (Empresa Brasileira de Pesquisa Agropecuária), classifica o uso e cobertura da Terra das áreas de desmatamento obtidas pelo PRODES. O objetivo é investigar sobre a dinâmica do desmatamento na região da Amazônia Legal, ou seja, investigar para qual finalidade as áreas são desmatadas com o intuito de obter um melhor entendimento do uso e cobertura da Terra nesta região [Almeida et al. 2016]. Os dados vetoriais do TerraClass são disponibilizados com uma frequência bienal. O INPE também desenvolve o Programa Queimadas que tem como objetivo o monitoramento de focos de queimadas e de incêndios florestais [INPE 2019b]. São produzidos dados pontuais com o atributo de data associado, com uma resolução temporal de quinze minutos. A Figura 2 apresenta uma visualização dos dados produzidos pelos programas citados.

4. Experimentos e resultados

O objetivo do experimento é ter uma visão inicial das funcionalidades disponíveis nas tecnologias GeoSpark, SpatialHadoop e ST-Hadoop, para realizar o processamento dos dados citados na seção 3. Com a finalidade de avaliar esse cenário, para a fase atual

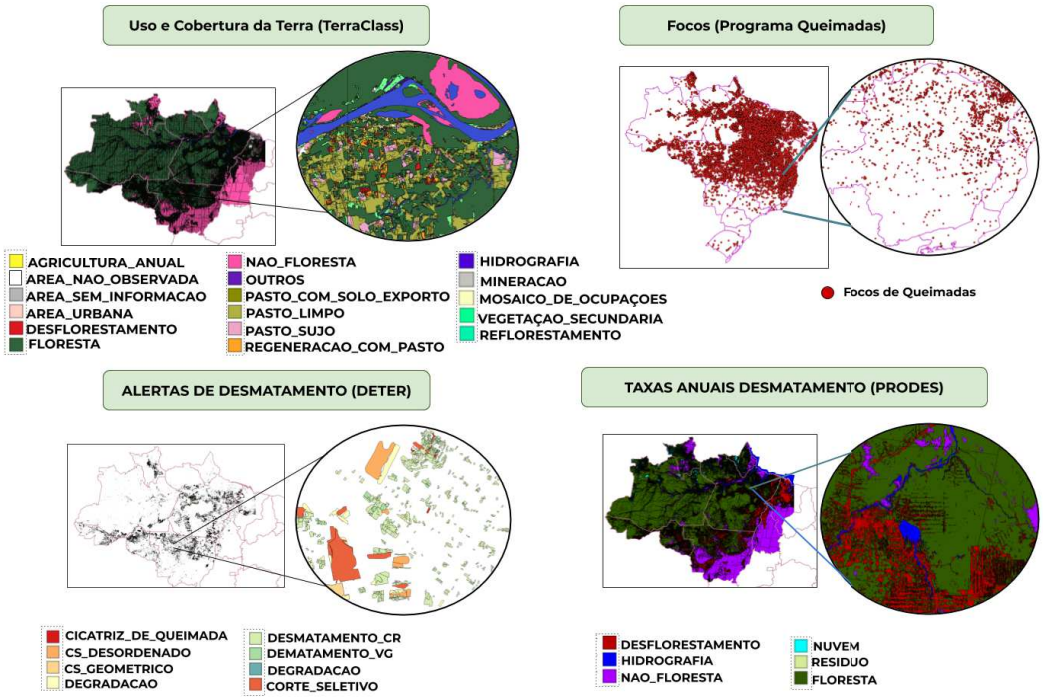

Figura 2. Programas de Monitoramento Ambietal do INPE

do trabalho, foi configurado ambiente com três *hosts*, utilizando as ferramentas Docker Machine[8] e Docker Swarm[9].

Em um primeiro caso, foram utilizados os dados de focos de queimadas do Brasil referente ao intervalo temporal dos anos de 2007 até 2018, disponibilizados pelo Programa Queimadas. Os dados foram armazenados no HDFS (*Hadoop Distributed File System*), e indexados pelas respectivas extensões com o índice *Grid*. Desta forma as operações realizadas nas diferentes tecnologias exploradas são feitas nos dados indexados. O experimento consistiu na execução de duas operações: `k-NN` e `Range`.

- Para a operação de `k-NN` busca-se responder a seguinte questão: Quais são os k focos de queimadas mais próximos ao um ponto P dado um intervalo temporal T.
- Para operação de `Range` busca-se responder: Dada uma geometria A, retornar o conjunto de dados de Q que interceptam A no intervalo temporal T.

As extensões SpatialHadoop e Geospark não oferecem operações nativas com o atributo temporal. Desta forma, foi necessário uma filtragem dos dados com base no atributo associado ao tempo, utilizando a linguagem de alto nível Pigeon[10]. Na extensão ST-Hadoop é fornecida como parâmetro a granularidade de tempo (diária, mensal ou anual) em que deseja-se realizar as operações. A Figura 3 apresenta o resultado da consulta `k-NN` na extensão GeoSpark. São apresentados os mil focos de queimadas mais próximos ao um ponto P de coordenadas $(x : -61.13, y : -4.52)$ dado um intervalo temporal `T:[2017-12-30, 2017-12-31]`. A Figura 4 resultado da consulta `Range` na

[8]https://docs.docker.com/machine/

[9]https://docs.docker.com/swarm/provision-with-machine/

[10]https://github.com/aseldawy/pigeon

extensão ST-Hadoop para a geometria A de coordenadas `x_1:-58.14,y_1:-14.51` e `x_2:-54.51,y_2:-11.06` para o ano de 2010 no mês de maio.

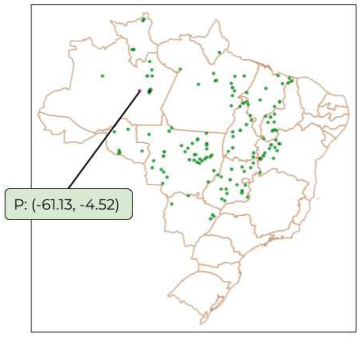

Figura 3. Resultado consulta k-NN GeoSpark

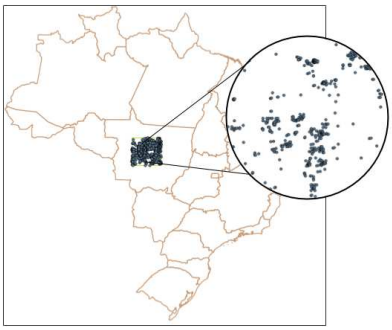

Figura 4. Resultado consulta Range no ST-Hadoop

5. Considerações Finais

Com base nesses experimentos, concluímos que:

- `Dados de entrada:` O Geospark apresenta um suporte a diversos formatos de dados de entrada, como Shapefiles e GeoJSON. Enquanto que SpatialHadoop e ST-Hadoop suporta apenas o formato CSV.
- `Tipos de dados:` Geospark e SpatialHadoop possuem suporte para os tipos de dados espaciais Ponto, Linha e Polígono. Apesar do ST-Hadoop dar suporte nativo para dados espaço-temporais, ele possui apenas o tipo de dado `ST_Point` para representar dados espaço-temporais. Esse tipo é uma tupla composta pela localização (x, y) e o tempo associado.
- `Operações:` Geospark e SpatialHadoop não trabalham diretamente com operações espaço-temporais. Deve-ser utilizar uma linguagem de alto nível para processar os dados que possuem o tempo como atributo.

Baseado nesses experimentos podemos concluir que as extensões analisadas, GeoSpark, SpatialHadoop e ST-Hadoop, não possuem todos os tipos de dados e operações espaço-temporais de forma nativa para atender todas as demandas de processamento dos dados espaço-temporais produzidos pelos programas de monitoramento do INPE. Por exemplo, nenhuma delas fornece um tipo de dado espaço-temporal para representar os polígonos de alertas de desmatamento que possuem tempos de observação associados. Além disso, nenhuma dessas extensões é capaz de executar uma junção espaço-temporal para realizar um cruzamento entre os focos de queimadas com os polígonos de alertas de desmatamento do DETER.

Portanto, seria necessário um grande esforço de programação para estender essas extensões com novos tipos de dados, operações e estruturas de índices espaço-temporais para atender todas as necessidades de processamento dos programas de monitoramento do INPE. Apesar dessa dificuldade, adicionar novas operações e tipos de dados de forma nativa a ferramentas consolidadas como Hadoop e Spark se mostra uma vertente promissora. Como trabalho futuro, pretende-se explorar a adição de uma estrutura de indexação

com suporte espaço-temporal para os tipos de dados Ponto, Linha e Polígono em uma das ferramentas abordadas no trabalho.

Agradecimentos

O presente trabalho foi realizado com apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil.

Referências

Aguiar, D. A., Silva, W. F., Rudorff, B. F., and Silva, J. S. (2010). MODIS Time Series to Assess Pasture Land. In *2010 IEEE International Geoscience and Remote Sensing Symposium*.

Alam, M. M., Ray, S., and Bhavsar, V. C. (2018). A Performance Study of Big Spatial Data Systems. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, pages 1–9. ACM.

Almeida, C. A. d., Coutinho, A. C., Esquerdo, J. C. D. M., et al. (2016). High spatial resolution land use and land cover mapping of the Brazilian Legal Amazon in 2008 using Landsat-5/TM and MODIS data. *Acta Amazonica*, 46:291 – 302.

Arvor, D., Simoes, M., Dubreuil, V., et al. (2011). Analyzing the agricultural transition in Mato Grosso, Brazil, using satellite-derived indices. *Applied Geography*, 32:702–713.

CEOS (2018). *The Earth Observation Handbook-Satellite Earth Observations in Support of the Sustainable Development Goals*.

Diniz, C. G., de Almeida Souza, A. A., Santos, D. C., et al. (2015). DETER-B: The New Amazon Near Real-Time Deforestation Detection System. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(7):3619–3628.

Foley, J. A., DeFries, R., Asner, G. P., et al. (2005). Global Consequences of Land Use. *Science (New York, N.Y.)*, 309:570–4.

García-García, F., Corral, A., Iribarne, L., et al. (2017). A Comparison of Distributed Spatial Data Management Systems for Processing Distance Join Queries. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pages 214–228, Cham. Springer International Publishing.

Gómez, C., White, J. C., and 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.

INPE (2019a). Monitoramento da floresta amazônica brasileira por satélite. `http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes`. Acesso: 19/08/2019.

INPE (2019b). Portal do monitoramento de queimadas e incêndios. `http://www.inpe.br/queimadas`. Acesso: 20/09/2019.

Lenka, R., Barik, D. R., Gupta, N., et al. (2016). Comparative Analysis of Spatialhadoop and Geospark for Geospatial Big Data Analytics. *2nd International Conference on Contemporary Computing and Informatics (IC3I 2016)*.

Pandey, V., Kipf, A., Neumann, T., and Kemper, A. (2018). How Good Are Modern Spatial Analytics Systems? *Proc. VLDB Endow.*, 11(11):1661–1673.