Avaliação Comparativa de Estratégias de Particionamento para Dados Raster em Bancos de Dados Multidimensionais

Marco Túlio Alves de Barros[1], **Daniel dos Santos Kaster**[1]

[1]Departamento de Computação – Universidade Estadual de Londrina (UEL)
Caixa Postal 10.011 – CEP 86057-970 – Londrina – PR – Brasil

marcotulio.barros@uel.br, dskaster@uel.br

Abstract. *This project aims to conduct a quantitative experiment to evaluate raster data partitioning methods in multidimensional databases. Given the increasing need for efficient manipulation of large volumes of geospatial data, whether for feature extraction in machine learning or other applications, this research is essential for identifying the most effective techniques in terms of response time and storage and query efficiency. The objectives include defining performance metrics, such as recovery time and space usage, and testing different partitioning approaches in real-world contexts. The proposed methods involve a critical literature review, analysis of tools like GDAL, PostGIS, and RasDaMan, and implementing benchmarks on a representative dataset. The expected outcomes of this research will provide valuable insights to optimize and scale current implementations, contributing to more efficient use of available tools and improving productivity in practical applications.*

Resumo. *Este projeto visa conduzir um experimento quantitativo para avaliar os métodos de particionamento de dados raster em bancos de dados multidimensionais. Dada a crescente necessidade de manipulação eficiente de grandes volumes de dados geoespaciais, seja extração de atributos para aprendizado de máquina ou outras aplicações, esta pesquisa é crucial para identificar as técnicas mais eficazes em termos de tempo de resposta e eficiência de armazenamento e consulta. Os objetivos incluem definir métricas de desempenho, como tempo de recuperação e uso de espaço, e testar diferentes abordagens de particionamento em contextos reais. Os métodos propostos envolvem uma revisão crítica da literatura, a análise de ferramentas como GDAL, PostGIS e RasDaMan, e a implementação de benchmarks em um conjunto de dados representativo. Espera-se que os resultados desta pesquisa forneçam insights valiosos para otimizar e escalar implementações atuais, contribuindo para um uso mais eficiente das ferramentas disponíveis e melhorando a produtividade em aplicações práticas.*

1. Introdução

Bancos de Dados são ferramentas robustas e consolidadas, com vertentes especializadas em diversas subáreas. O avanço nas tecnologias de sensoriamento remoto traz inúmeros benefícios para a sociedade, permitindo a obtenção de recursos preciosos em diversos contextos, como na agropecuária [10], [18]; previsões meteorológicas [11]; posicionamento de torres de celular [4]. No entanto, o tratamento de dados complexos, como dados geográficos de raster, que são basicamente vetores multidimensionais, envolve um volume

massivo de informações [7], [20]. Esse desafio exige o desenvolvimento e a aplicação de metodologias de particionamento para otimização de armazenamento e consulta de dados.

Neste cenário, destacam-se pesquisadores como Peter Baumann, que se aprofundou em bancos de dados de vetores e particionamento [1], [2], [5], [9], [19]; Ramon Antonio Rodriges Zalipynis, que investigou as ferramentas disponíveis na área [21]; e Michael Stonebraker, cujos artigos abordam arquiteturas de dados persistentes [15]. Apesar da existência de soluções robustas e pesquisas significativas, a área ainda carece de detalhamento e explicações claras dos parâmetros envolvidos.

Na literatura e setor empresarial são descritas algumas possibilidades, tal como a abordagem mais trivial é a persistência do arquivo. Neste caso um dado raster, como um todo, exigindo manipulá-lo por inteiro e acarretando em grande desperdício de recursos. Outra abordagem, que apresenta uma evolução da anterior, é a recuperação de dados *in situ*, utilizando ferramentas auxiliares, que permitem a recuperação e análise das informações por faixas ou regiões, reduzindo a necessidade de movimentação de grandes volumes de dados [15].

Avançando em relação à abordagem trivial, uma da principais técnicas em grandes bancos de dados geoespaciais é o particionamento de dados raster seguindo algumas estratégias. A mais comum delas é o particionamento regular, que permite a divisão em pedaços menores e idênticos ao original, podendo otimizar o armazenamento (pedaços que ocupam espaço em disco múltiplo do tamanho do bloco). Embora essa solução apresente bons resultados, ela pode gerar desperdício de recursos na recuperação dos dados, pois, independentemente do tamanho, a área de interesse contida no bloco será integralmente recuperada [15], [9] [5].

Um dos problemas em aberto na literatura é na recuperação de unidades singulares ou formatos atípicos em dados raster de forma automática ou que adapte às situações, visto que os particionamentos estão regidos por parâmetros como tamanho de bloco, dimensões ou características específicas dos dados. Dessa forma, existem artigos [5], [19] que fizeram análises semelhantes, mas deixaram lucunas em relação aos parâmetros variados, experimentos conduzidos, tipos de dados utilizados e conclusões métricas exatas.

Neste contexto, a principal contribuição deste trabalho é abordar essas lacunas, fornecendo uma análise detalhada e comparativa de diferentes estratégias de particionamento de dados raster. Ao realizar uma avaliação meticulosa e quantitativa dessas técnicas, este estudo busca identificar as abordagens mais eficientes, seja em tempo, espaço ou outro atributo proposto, para a recuperação de dados raster, oferecendo insights valiosos para a escolha de estratégias de armazenamento e, principalmente, consulta.

As novas tecnologias de disco e algoritmos de particionamento, juntamente com a necessidade de manipular grandes volumes de dados, tornam este estudo especialmente relevante. Sistemas como RasDaMan oferecem capacidades avançadas, mas sua complexidade pode limitar sua adoção. Por outro lado, ferramentas populares como GDAL [6] e Python Pandas [12] utilizam abordagens triviais ou *in situ*. Assim, uma pesquisa sobre a eficiência e praticidade dessas ferramentas pode melhorar significativamente a produtividade. Além disso, soluções comuns de aprendizado de máquina frequentemente manipulam dados multidimensionais *in situ* [13], [8], e implementar abordagens integradas e eficientes para armazenamento e recuperação de dados pode potencializar a escalabilidade

de projetos organizados dessa forma.

O objetivo principal deste projeto é conduzir um experimento, destacando aplicabilidade em situações reais, para medir e testar tamanhos de blocos adequados, avaliar o impacto efetivo da distribuição de datasets em diferentes programas e algoritmos. E consequentemente propor e estudar os principais parâmetros que influenciam a performance, como tamanho de bloco, podendo ainda explorar outras técnicas.

Por fim, visa-se um enfoque em análises e medidas contextualizadas às tecnologias atuais. As conclusões fundamentadas em dados robustos e estratégias de métricas bem implementadas proporcionarão uma base sólida para decisões informadas. Isso permitirá a escolha de sistemas e abordagens mais adequadas para diferentes cenários de manipulação de dados geoespaciais. Medidas quantitativas de impacto são extremamente valiosas e devem permitir que usuários e projetistas que trabalham com dados geográficos tomem decisões embasadas e esclarecedoras, otimizando o uso dos recursos disponíveis.

Neste projeto, a Seção 2 apresenta a fundamentação teórico-metodológica e o estado da arte, a Seção 3 irá abordar os objetivo gerais da pesquisa, a Seção 4 transcorre a respeito das etapas propostas para atingir os objetivos, a Seção 5 inclui o cronograma estipulado inicialmente para cada atividade proposta e, por fim, a Seção 6 aborda as contribuições e o resultados esperados após a conclusão dessa pesquisa.

2. Fundamentação Teórico-Metodológica e Estado da Arte

Esta seção explora os trabalho correlatos, conceitos, métodos e técnicas essenciais para a compreensão do particionamento de raster em bancos de dados multidimensionais, proporcionando a base teórica necessária para analisar soluções existentes, desenvolver novas abordagens e realizar comparações quantitativas. Essa base teórica incluirá referências de livros, teses e dissertações, fornecendo detalhamento necessário para analisar soluções existentes e desenvolver novas abordagens. Além disso, identificar os conceitos independentemente do propósito de uso, preparando o terreno para a revisão do estado da arte.

2.1. Trabalhos Correlatos

A crescente demanda por dados geoespaciais de alta precisão e volume tem impulsionado significativas pesquisas no campo de particionamento de dados raster. No contexto da agropecuária, [10] e [18] exploraram o uso de dados de sensoriamento remoto para otimizar a produtividade agrícola, evidenciando a importância de soluções eficientes para manipulação de grandes volumes de dados. No mesmo sentido, [11] destacou a aplicação desses dados em previsões meteorológicas, reforçando a relevância de técnicas eficazes de armazenamento e consulta.

Os desafios associados ao tratamento de dados raster são amplamente discutidos na literatura. Estudos de pesquisadores como Peter Baumann [1], [2], que em seus trabalhos sobre bancos de dados de vetores e particionamento, desenvolveu uma base teórica robusta para o armazenamento eficiente de dados multidimensionais. Baumann trabalhou em conjunto com Paula Furtado e Norbert Widmann expandindo essas pesquisas [5], [19].

Além de Baumann, Ramon Antonio Rodriges Zalipynis [21] examinou as ferramentas disponíveis na área, como GDAL e RasDaMan, oferecendo uma visão crítica sobre suas capacidades e limitações, além de compilar alguns detalhes importantes para

estudo da área. Por outro lado, Stonebraker [15], pesquisador de grande influência na área, concentrou-se em arquiteturas de dados persistentes, contribuindo para a compreensão das melhores práticas em armazenamento de grandes volumes de dados geoespaciais.

Na pesquisa [5], em especial, apesar de antiga é fundamental para o estudo, agrupando os principais conceitos e expandindo os testes de particionamento. É a mais próxima da proposta desse projeto, porém as variações incluem o tamanho dos *chunks*, áreas de interesse e tipos de consultas, sem fornecer muitos detalhes e especificações acerca deles. Outras pesquisas seguem a mesma linha [19], [9] [3] explorando consultas e tamanhos sem fornecer específicos, como variações exatas em blocos e *chunks* ou características bem definidas dos conjuntos testados, por exemplo.

Apesar dessas contribuições importantes e da formação de uma base robusta, a literatura ainda carece de detalhamentos sobre parâmetros experimentais e variações de dados evidenciado dificultando a replicação e ampliação de estudos na comunidade acadêmica. Assim, nessa carência, aliada com o certo intervalo entre as pesquisas mais relevantes e a evolução de tecnologias, é o contexto em que esse projeto se insere.

2.2. Conceitos Fundamentais

Dados raster são representações de informações espaciais em forma de grades ou matrizes. Cada célula dessas grades contém um valor específico associado a uma localização geográfica precisa. Conforme definido por Baumann [2], esses dados são comumente utilizados em imagens, representações de mapas, dados de satélites e outras aplicações geoespaciais. A Figura 1 ilustra como ocorre, didaticamente, essa representação real para raster.

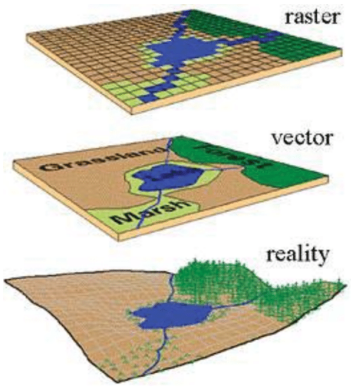

Figura 1. Fonte [17]. Ilustração didática para mostrar como dados reais são representados em raster, neste caso, uma matriz (vetor bidimensional) com todos as partes iguais.

Devido ao grande volume de informações raster, especialmente em contextos geoespaciais, torna-se crucial adotar uma organização otimizada para armazenamento e consultas, visando reduzir tempos de consultas e reduzir desperdício de recursos computacionais. Ambientes que trabalham com essas informações, por exemplo, fazem consultas simples mas com retorno custoso, logo o impacto de performance é perceptível ao usuário.

Para otimizar a organização dos dados, é essencial considerar quatro principais técnicas de particionamento, ilustradas na Figura 2. Essas técnicas abordam diferentes

padrões de *chunking*[1], e suas aplicações específicas. Essas técnicas influenciam diretamente a eficiência do armazenamento em disco e a recuperação de dados, como será detalhado a seguir.

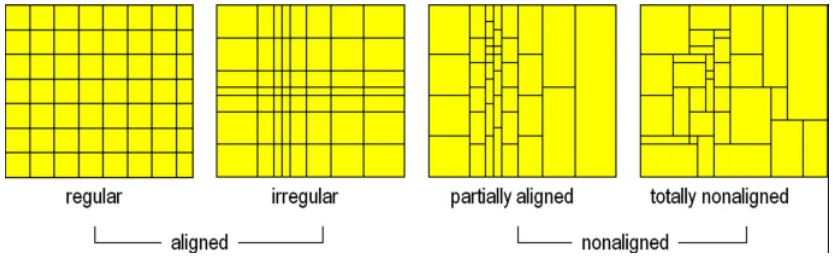

Figura 2. Fonte [2]. Diferentes padrões de chunks

- **Regular Alinhado**: Neste método, os dados são divididos em blocos de tamanhos uniformes e alinhados em uma grade regular. A vantagem principal desse método é a simplicidade e a previsibilidade no acesso aos dados. Como os blocos são uniformes e alinhados, é fácil calcular a localização de qualquer dado específico. Isso minimiza a fragmentação e pode melhorar a eficiência de leitura e escrita em sistemas de arquivos que se beneficiam de acessos sequenciais e regulares [21], [2], [5], [15], [14].
- **Irregular Alinhado**: Aqui, os blocos também são alinhados, mas podem ter tamanhos variados. A grade é regular, mas os blocos podem ter dimensões diferentes. Este método permite uma maior flexibilidade na adaptação aos dados, o que pode reduzir a quantidade de espaço desperdiçado. No entanto, a variabilidade dos tamanhos dos blocos pode complicar o cálculo da localização dos dados e potencialmente aumentar a fragmentação, especialmente em sistemas de arquivos que não lidam bem com tamanhos de bloco variáveis [21], [2], [5], [15], [14].
- **Parcialmente Alinhado**: Nesta abordagem, alguns blocos são alinhados enquanto outros não. Isso pode ocorrer devido à necessidade de ajustar os blocos às bordas irregulares dos dados. Esta técnica pode ser útil quando se trabalha com dados que têm áreas de interesse específicas que não se encaixam perfeitamente em uma grade regular. Embora possa aumentar a eficiência do uso do espaço em certas situações, pode também introduzir complexidade adicional na recuperação de dados e no gerenciamento de blocos desalinhados [21], [2], [5], [15], [14].
- **Totalmente Desalinhado**: Neste método, os blocos são de tamanhos e formas variadas e não seguem uma grade regular. Este é o método mais flexível e pode ser otimizado para minimizar o desperdício de espaço e maximizar a eficiência da recuperação de dados para casos de uso específicos. No entanto, essa flexibilidade vem ao custo de uma maior complexidade no cálculo das localizações dos dados e potencialmente maior fragmentação, o que pode impactar negativamente a performance de leitura e escrita [21], [2], [5], [15], [14].

Além disso, é possível combinar as estratégias anteriores com situações em que, devido a organização do armazenamento, consultas e cargas são altamente prejudicadas, trazendo dados desnecessários para tal, como no exemplo 3. Nele busca evidenciar como consultas pelo conjunto rachurado acarreta na recuperação de muitos outros dados desnecessários, aumentando tempo e recursos computacionais para tal.

[1]Outros termos comuns na literatura são *tiles* e *tiling* [14]

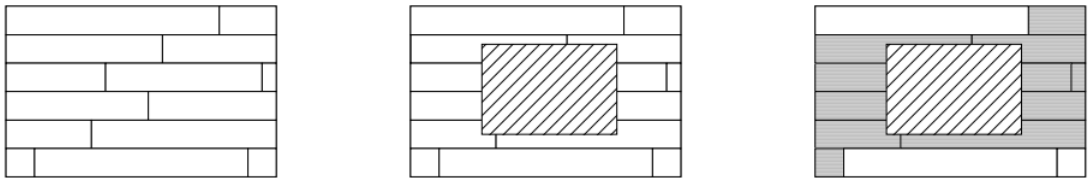

Figura 3. Fonte [9]. Ilustração de como as consultas são afetadas pelos particionamentos

A figura 4 contém um *datacube* com particionamento irregular alinhado (faces frontal e direita) e parcialmente alinhado (face superior). Pode representar, por exemplo, um acumulado diário em uma determinada região geográfica. Esta representação visual ajuda a entender como diferentes métodos de particionamento podem ser aplicados e mesclados para otimizar o armazenamento e a recuperação de dados em sistemas multidimensionais.

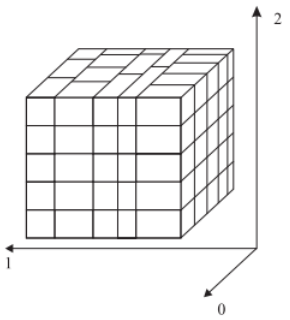

Figura 4. Fonte [5]. Exemplo do *datacube*

2.3. Ferramentas para Armazenar ou Manipular Dados Raster

Para aplicar os conceitos vistos da anteriormente e gerar funcionalidades práticas, diversas ferramentas estão disponíveis para explorar as possibilidades de particionamento e armazenamento de dados raster, oferecendo diferentes abordagens e capacidades. A seguir, detalhamos algumas das principais ferramentas:

- GDAL(Geospatial Data Abstraction Library)[2] é uma biblioteca de código aberto fundamental para o processamento de dados geoespaciais. Ela oferece suporte para a leitura e escrita de diversos formatos de dados raster e vetoriais. Com ele é possível realizar operações como reamostragem, transformação de projeções, recorte e fusão de imagens, bem como manipulação direta dos dados em memória. É amplamente utilizada devido à sua flexibilidade e ampla compatibilidade com diferentes formatos de dados geoespaciais. No contexto do particionamento de dados raster, o GDAL permite a manipulação local dos dados brutos, o que pode ser útil para experimentação e testes iniciais.
- PostgreSQL[3] é um banco de dados relacional de código aberto amplamente utilizado, conhecido por sua robustez e extensibilidade. A extensão PostGIS[4] adiciona

[2]Disponível em https://gdal.org/
[3]Disponível em https://www.postgresql.org/
[4]Disponível em https://postgis.net/

suporte a dados geoespaciais, transformando-o em um poderoso sistema de gerenciamento de banco de dados geoespacial. O PostGIS permite o armazenamento, consulta e análise de dados geoespaciais de maneira eficiente, utilizando funções e operadores específicos para dados espaciais. No contexto do particionamento de dados raster, o PostgreSQL + PostGIS oferece ferramentas para o armazenamento e recuperação eficientes, além de suporte para indexação espacial, que pode melhorar significativamente o desempenho das consultas geoespaciais.

- RasDaMan[5] é um sistema de gerenciamento de banco de dados especializado em dados geoespaciais e multidimensionais. Ele é projetado para lidar com grandes volumes de dados raster, permitindo operações complexas diretamente no banco de dados. Suporta consultas em várias dimensões, operações de álgebra de arrays e oferece otimizações específicas para o particionamento e indexação de dados raster. Este sistema é ideal para aplicações que requerem alta performance e escalabilidade no gerenciamento de grandes conjuntos de dados geoespaciais. Por ser extremamente complexo apenas especialistas costumam lidar com ele, porém é esperado que apresente a melhor performance.

2.4. Técnicas de Particionamento, Armazenamento e Consulta

Após a apresentação dos principais conceitos e ferramentas, neste projeto os produtos dos particionamentos serão referidos como *chunks*. A forma como os raster são internamente armazenados, em relação à separação dos *chunks*, é um fator determinante para a eficiência do sistema. A escolha entre particionamento regular ou irregular deve considerar as características específicas de cada problema [5]. Além disso, questões técnicas de hardware, como o tamanho dos blocos de dados e a velocidade de leitura da memória terciária, também desempenham um papel crucial [3].

1. Particionamento Regular e Irregular: como descritos anteriormente com a figura 2, o impacto de cada técnica pode variar, por isso pretende-se medir e avaliar os principais fatores que podem otimizar a performance proposta.
2. Hardware: o tamanho dos blocos de dados, por exemplo, deve ser cuidadosamente ajustado para maximizar a velocidade de leitura e escrita na memória terciária. Blocos menores podem aumentar a granularidade do acesso aos dados, mas também podem aumentar o overhead de gerenciamento. Blocos maiores, por outro lado, podem reduzir o overhead, mas podem não ser tão eficientes em termos de acesso a pequenas áreas de interesse.

Com o progresso do projeto, novos elementos podem ser considerados e explorados.

3. Objetivos

3.1. Objetivo Geral

O principal objetivo deste projeto é realizar uma análise comparativa quantitativa de diferentes estratégias de particionamento para armazenamento de dados raster. Visa-se determinar a abordagem que minimiza a recuperação de blocos de uma mesma consulta, reduz o desperdício de recursos computacionais e minimiza o tempo de resposta das consultas. Além disso, o projeto busca aplicar medidas comparativas para avaliar o impacto de diferentes ferramentas e abordagens no acesso aos dados, espaço de armazenamento e tempo de execução.

[5]Disponível em http://www.rasdaman.org/

3.2. Objetivos Específicos

- Identificar e descrever estratégias de particionamento, com vantagens e limitações.
- Implementação de estratégias selecionadas em ambiente controlado e fazendo uso das ferramentas propostas.
- Definir os operadores que serão utilizados, como, por exemplo, redução temporal (acumulados) e redução espacial (médias).
- Aplicar comparações que incluirão tempo de recuperação de dados, espaço de armazenamento utilizado e tempo de execução das consultas e devem ser capazes de capturar as nuances das diferentes estratégias.
- Análises de desempenho conduzindo testes e buscando identificar padrões de eficiência e ineficiência em cada abordagem.

4. Procedimentos metodológicos/Métodos e técnicas

A Figura 5 ilustra brevemente as etapas propostas e o fluxo de execução geral, e em seguida estão descritas as atividades previstas do projeto, como mais detalhes.

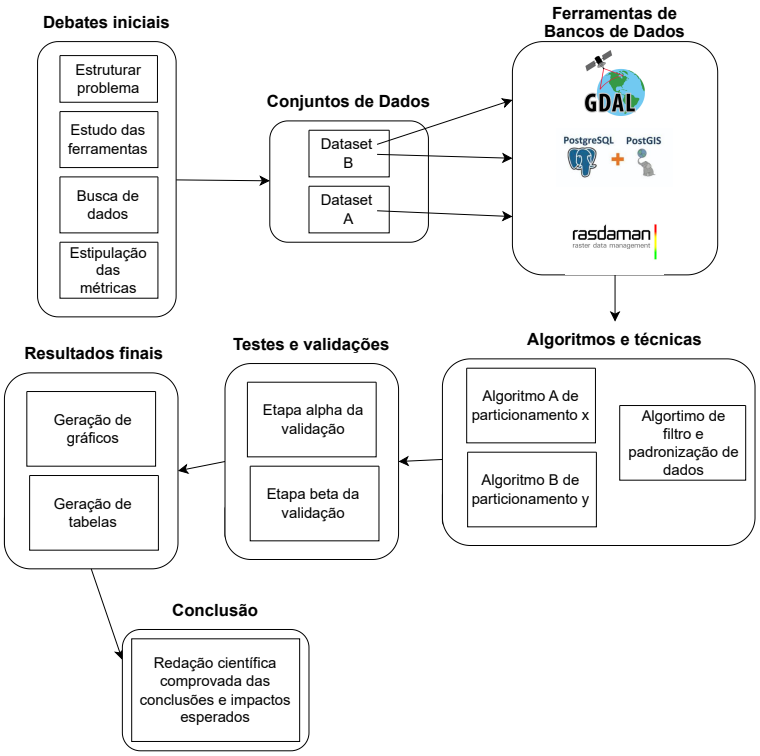

Figura 5. Fluxograma das atividades de pesquisa planejada.

1. **Revisão bibliográfica:** Considerando que o campo de particionamento de dados raster em bancos de dados multidimensionais é uma área bem estabelecida com vasta literatura, é crucial definir como os conceitos tradicionais serão atualizados e aplicados em contextos modernos. Isso envolve a revisão crítica de pesquisas anteriores, avaliar métodos descritos e testar as implementações proposta, averiguar quais apresentam melhores resultados e que poderão contribuir para aplicações

futuras, identificação de lacunas e a formulação de uma abordagem inovadora que contribua de maneira significativa para o avanço do estado da arte atual.

2. **Análise das abordagens existentes:** Esta etapa inicial envolve uma análise detalhada dos sistemas e ferramentas existentes para particionamento de dados raster, como GDAL, PostGIS e RasDaMan. Será produzido um relatório abrangente que lista as principais características, pontos fortes e limitações de cada sistema, destacando aspectos relevantes para o projeto, como eficiência de armazenamento e tempo de recuperação de dados.
3. **Estabelecimento de um conjunto de dados de teste:** Selecionar e preparar um conjunto de dados que represente realisticamente os desafios enfrentados no particionamento de dados raster. Isso inclui dados geoespaciais de diferentes fontes, como imagens de satélite e mapas digitais. O resultado será um conjunto de dados diversificado e representativo, pronto para ser utilizado nas análises e testes subsequentes. As fontes podem incluir o MapBiomas[6][16] e o Portal de Mapas do IBGE[7].
4. **Modelagem e discussão das análises quantitativas:** Propor e documentar métricas de benchmark que serão utilizadas para avaliar a eficiência das estratégias de particionamento. Isso envolve a pesquisa na literatura para identificar métricas existentes, como tempo de resposta, uso de espaço de armazenamento. O resultado será um conjunto bem definido de métricas de benchmark prontas para implementação.
5. **Modelagem e discussão dos parâmetros que serão variados e observados:** Definir os parâmetros específicos que serão ajustados e observados durante os testes, como o tamanho dos *chunks*, a estrutura de indexação, e a configuração do hardware utilizado. Esta etapa exige uma revisão contínua e ajuste dos parâmetros com base em resultados empíricos e na literatura existente. O resultado será um plano detalhado dos parâmetros a serem estudados, juntamente com a justificativa científica para cada escolha.
6. **Implementação:** Programar e executar os sistemas de particionamento utilizando as métricas definidas. Esta etapa envolve a implementação de algoritmos, configuração dos ambientes de teste, execução das estratégias de particionamento com diferentes parâmetros e coleta de dados estatísticos. O resultado esperado é um conjunto de scripts e programas que aplicam as métricas de benchmark de forma consistente e justa.
7. **Execução:** Analisar os dados coletados, gerando gráficos, tabelas e relatórios que comparem as diferentes estratégias de particionamento. Esta etapa finaliza a parte prática do projeto e visa fornecer uma visualização clara dos resultados, destacando as abordagens mais eficientes. O resultado será uma série de relatórios detalhados e visualizações gráficas que elucidam os pontos fortes e fracos de cada estratégia.
8. **Escrita final do texto:** Documentar todo o processo, desde a revisão da literatura até os resultados dos testes e a análise comparativa. Esta etapa será intercalada com as atividades práticas, permitindo que os resultados sejam formalizados e integrados no texto à medida que são obtidos. O produto final será o trabalho de conclusão do curso completo, pronto para submissão e apresentação.

[6] Acessível em: https://brasil.mapbiomas.org/mapas-de-referencia/

[7] Acessível em: https://portaldemapas.ibge.gov.br/portal.php

5. Cronograma de Execução

Essa seção visa identificar o período estimado para cada atividade relacionada em 4, além de breve justificativa para tal.

Atividades:

1. Organizar as ideias e revisão bibliográfica;
2. Analisar das abordagens existentes;
3. Definir e obter conjuntos de teste;
4. Discutir e estabelecer análises quantitativas;
5. Discutir possíveis parâmetros a serem variados;
6. Implementar, ou esquematizar, os experimentos;
7. Executar e analisar testes;
8. Documentar todas as etapas.

Tabela 1. Cronograma de Execução

	jun	jul	ago	set	out	nov	dez	jan	fev	mar
Atividade 1	x	x								
Atividade 2		x	x							
Atividade 3		x	x							
Atividade 4		x	x							
Atividade 5			x	x	x	x				
Atividade 6				x	x	x	x			
Atividade 7							x	x	x	
Atividade 8	x	x	x	x	x	x	x	x	x	

6. Contribuições e/ou Resultados Esperados

A principal contribuição científica deste projeto será a condução de um conjunto robusto de análises quantitativas para avaliar a eficiência de diferentes estratégias de particionamento de dados raster. Ao fim deste projeto, teremos dados confiáveis que apontem qual abordagem é mais eficiente em determinadas situações, além de um conjunto de parâmetros explorados em detalhes e as definições dos conjuntos de dados utilizados.

Essas métricas e análises permitirão que pesquisadores, desenvolvedores e usuários finais tomem decisões mais informadas sobre as estratégias de particionamento e armazenamento de dados raster em casos de uso específicos, tornando operações mais escaláveis e eficientes. Portanto, espera-se que, ao encerrar, seja possível evidenciar as abordagens atuais e compará-las com as abordagens ideais, que tendem a otimizar processos e preencher as lacunas existentes na literatura.

Além disso, espera-se que este projeto resulte na proposta de novas abordagens, como talvez um particionamento dinâmico e adaptativo, e otimizações para as ferramentas existentes, com base nas análises realizadas. A aplicação dessas novas estratégias poderá levar a melhorias significativas na eficiência e desempenho dos sistemas de armazenamento e recuperação de dados raster, contribuindo assim para o avanço do estado da arte nessa área, sendo relevante para a comunidade científica e indústria.

7. Espaço para assinaturas

Londrina, *29 de julho de 2024*.

Aluno

Orientador

Referências

[1] Peter Baumann, Andreas Dehmel, Paula Furtado, Roland Ritsch, and Norbert Widmann. The multidimensional database system rasdaman. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 575–577, 1998.

[2] Peter Baumann, Dimitar Misev, Vlad Merticariu, and Bang Pham Huu. Array databases: Concepts, standards, implementations. *Journal of Big Data*, 8:1–61, 2021.

[3] Philippe Cudre-Mauroux, Hideaki Kimura, Kian-Tat Lim, Jennie Rogers, Samuel Madden, Michael Stonebraker, Stanley B Zdonik, and Paul G Brown. Ss-db: A standard science dbms benchmark. *Under submission*, 114, 2010.

[4] Esri. Mapa topográfico mundial. ArcGIS Online, 19 de fevereiro de 2012. Escala Não Fornecida.

[5] Paula Furtado and Peter Baumann. Storage of multidimensional arrays based on arbitrary tiling. In *Proceedings 15th International Conference on Data Engineering (Cat. No. 99CB36337)*, pages 480–489. IEEE, 1999.

[6] GDAL/OGR contributors. *GDAL/OGR Geospatial Data Abstraction software Library*. Open Source Geospatial Foundation, 2024.

[7] Fei Hu, Mengchao Xu, Jingchao Yang, Yanshou Liang, Kejin Cui, Michael M. Little, Christopher S. Lynnes, Daniel Q. Duffy, and Chaowei Yang. Evaluating the open source data containers for handling big geospatial raster data. *ISPRS International Journal of Geo-Information*, 7(4), 2018.

[8] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs, 2021.

[9] Paolo Marques, Paula Furtado, and Peter Baumann. An efficient strategy for tiling multidimensional olap data cubes. pages 13–24, 01 1998.

[10] Kh T Murodilov, II Muminov, and RR Abdumalikov. Using geospatial data to optimize agricultural production in region/country. *Educational Research in Universal Sciences*, 2(4):115–117, 2023.

[11] Olavo Pereira do Nascimento. Técnicas de armazenamento e consulta de dados de precipitação obtidos por meio de imagens de radar. Trabalho de conclusão de curso, Universidade Estadual de Londrina, Londrina, 2023.

[12] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

[13] Mahbubur Rahman. A novel index-based multidimensional data organization model that enhances the predictability of the machine learning algorithms. In *Computer Science amp; Information Technology (CS amp; IT)*, MLNLP 2020. AIRCC Publishing Corporation, October 2020.

[14] Florin Rusu and Yu Cheng. A survey on array storage, query languages, and systems. *arXiv preprint arXiv:1302.0103*, 2013.

[15] Sunita Sarawagi and Michael Stonebraker. Efficient organization of large multidimensional arrays. In *Proceedings of 1994 IEEE 10th International conference on data engineering*, pages 328–336. IEEE, 1994.

[16] Carlos M. Souza, Julia Z. Shimbo, Marcos R. Rosa, Leandro L. Parente, Ane A. Alencar, Bernardo F. T. Rudorff, Heinrich Hasenack, Marcelo Matsumoto, Laerte G. Ferreira, Pedro W. M. Souza-Filho, Sergio W. de Oliveira, Washington F. Rocha, Antônio V. Fonseca, Camila B. Marques, Cesar G. Diniz, Diego Costa, Dyeden Monteiro, Eduardo R. Rosa, Eduardo Vélez-Martin, Eliseu J. Weber, Felipe E. B. Lenti, Fernando F. Paternost, Frans G. C. Pareyn, João V. Siqueira, José L. Viera, Luiz C. Ferreira Neto, Marciano M. Saraiva, Marcio H. Sales, Moises P. G. Salgado, Rodrigo Vasconcelos, Soltan Galano, Vinicius V. Mesquita, and Tasso Azevedo. Reconstructing three decades of land use and land cover changes in brazilian biomes with landsat archive and earth engine. *Remote Sensing*, 12(17), 2020.

[17] Marc Spiller and Claudia Agudelo. Mapping diversity of urban metabolic functions – a planning approach for circular urban metabolism. 01 2011.

[18] RC TAQUES and MAM ROCHA. Aptidão agrícola para a cultura da mamoneira (ricinus communis l.) no estado do espírito santo. In: CONGRESSO BRASILEIRO DE AGRONOMIA, 25., Vitória, ES.[Anais...] Vitória . . . , 2014.

[19] Norbert Widmann and Peter Baumann. Performance evaluation of multidimensional array storage techniques in databases. In *Proceedings. IDEAS'99. International Database Engineering and Applications Symposium (Cat. No. PR00265)*, pages 385–389. IEEE, 1999.

[20] Zhixin Yao, Jianqin Zhang, Taizeng Li, and Ying Ding. A trajectory big data storage model incorporating partitioning and spatio-temporal multidimensional hierarchical organization. *ISPRS International Journal of Geo-Information*, 11(12), 2022.

[21] Ramon Antonio Rodriges Zalipynis. Array dbms: past, present, and (near) future. *Proceedings of the VLDB Endowment*, 14(12):3186–3189, 2021.