Comparação de Desempenho na Indexação de *Big Geospatial Data* em Ambiente de Nuvem Computacional

João Bachiega Jr., Marco Sousa Reis, Maristela Holanda, Aletéia P. F. Araújo

[1]Departamento de Ciência da Computação – Universidade de Brasília (UNB)
Brasília – DF – Brasil

joao.bachiega.jr@gmail.com, ma@marcoreis.net,{mholanda, aleteia}@unb.br

Abstract. *With the growth of spatial data volume, known as Big Geospatial Data, some tools have been developed to allow the processing of this data in an efficient way, but for this it is fundamental to index the databases. The cloud computing has computational power and several other characteristics that are adherent to the execution of this type of application. This paper presents an analysis of indexing, operations and queries performed through SpatialHadoop in a test scenario provisioned in the cloud environment.*

Resumo. *Com crescimento do volume de dados espaciais, conceituado como Big Geospatial Data, algumas ferramentas foram desenvolvidas para permitir o processamento desses dados de forma eficiente, mas para isso é fundamental a indexação das bases de dados. A computação em nuvem possui poder computacional e diversas outras características que são aderentes para a execução deste tipo de aplicação. Este trabalho apresenta uma análise de indexações, operações e consultas realizadas através da ferramenta SpatialHadoop em um cenário de testes provisionado em ambiente de nuvem.*

1. Introdução

O enorme volume de dados geográficos gerados e disponibilizados nos últimos anos, conceituado como *Big GeoSpatial Data*, tem motivado pesquisadores a encontrarem uma solução para o processamento desses dados [Yang et al. 2017]. Ao mesmo tempo, tem-se a disponibilização de poder computacional capaz de suprir as necessidades geradas por estas aplicações, o que é encontrado na computação em nuvem, um modelo que possibilita acesso sob demanda a um vasto conjunto de recursos computacionais.

Para que as aplicações de *Big Geospatial Data* tenham um bom desempenho, uma tarefa importante é a indexação do conjunto de dados. No entanto, existem diferenças entre os métodos de indexação, fazendo com que a escolha do índice mais adequado ao conjunto de dados, às consultas e às operações a serem executadas, seja fundamental.

Este artigo propõe uma comparação de desempenho na indexação de *Big Geospatial Data* em ambiente de nuvem computacional, buscando indicar a configuração mais adequada para otimizar o desempenho (tempo) da aplicação, baseado nos tipos de dados contidos nos conjuntos de dados a serem processados, e nos parâmetros das consultas espaciais que serão realizadas.

Assim, este artigo está estruturado, em mais cinco seções. A Seção 2 apresenta o conceito de *Spatial Cloud Computing*. Na Seção 3 são apresentadas as características

do *SpatialHadoop*. Em seguida, os métodos para indexação são apresentados na Seção 4. Os trabalhos já desenvolvidos sobre este tema são detalhados na Seção 5. A Seção 6 apresenta os testes preliminares realizados. Por fim, a Seção 7 apresenta a conclusão do que foi analisado até o momento e os direcionamentos para a continuidade deste trabalho.

2. *Spatial Cloud Computing*

O termo *Big Geospatial Data* é um paradigma emergente para a grande quantidade de informações geográficas gerada, dada a crescente utilização de Sistemas de Informações Geográficas (SIG), atingindo *petabytes* de informações a cada dia [Eldawy and Mokbel 2015a]. Estes dados são gerados das mais diversas formas, tais como em mídias sociais, dispositivos móveis, satélites, entre outros. Além disso, esses dados têm sido gerados de uma maneira cada vez mais acelerada. O desafio de transformar grandes volumes de dados em conhecimento, exige requisitos de armazenamento, de acesso, de análise e de mineração dos dados.

A Computação em Nuvem é um modelo de entrega de poder computacional em forma de serviço que possui características que permitem o processamento de grandes volumes de dados, tais como a elasticidade para o provisionamento de recursos; o alto poder computacional obtidos através do compartilhamento de recursos; o amplo acesso que permite uma rápida comunicação; a obtenção de recursos de acordo com a demanda; e, por fim, a tarifação baseada apenas nos recursos que foram utilizados. Por todas estas características, a computação em nuvem mostra-se bastante aderente ao processamento de *Big Geospatial Data* [Li et al. 2010].

3. *SpatialHadoop*

O processamento de grandes volumes de dados espaciais tem demandado não apenas recursos computacionais robustos, mas também métodos eficientes. Nos últimos anos, diversas aplicações foram desenvolvidas utilizando os conceitos de *Hadoop* para otimizar o processamento desses dados, tais como: *GIS Tools on Hadoop* [Hoel and Park 2014] e *Hadoop-GIS* [Aji et al. 2013]. Em [Eldawy and Mokbel 2013] foi apresentado o *SpatialHadoop*, um *framework* que está incorporado no *Hadoop*, implementando as funcionalidades espaciais no seu interior e também utilizando índices espaciais. Desta forma, o *SpatialHadoop* tem mostrado desempenho superior quando comparado com todas as demais aplicações existentes até então.

O núcleo do *SpatialHadoop* consiste em quatro camadas, as quais são [Eldawy and Mokbel 2015b]: a Camada de Linguagem que utiliza o *Pigeon*, uma linguagem *SQL-like*, que suporta os tipos de dados padrões do *Open Geospatial Consortiums* (OGC); a Camada de Operações que encapsula a implementação de diversas operações espaciais que utilizam os índices espaciais; a Camada *MapReduce* que para ser capaz de lidar com arquivos indexados espacialmente, introduz dois novos componentes – *SpatialFileSplitter* e *SpatialRecordReader*; e, por fim, a Camada de Armazenamento que adiciona índices espaciais para superar uma limitação do *Hadoop*, que provê suporte apenas para arquivos não indexados do tipo *heap*, organizando o seu índice em níveis indexação globais e locais.

4. Indexação para *Big Geospatial Data*

O processamento de operações espaciais é fortemente influenciado pelo uso de estruturas de dados e algoritmos de pesquisa conhecidos como Métodos de Acesso Multidimensionais (MAM) [Gaede and Günther 1998]. Estes métodos são projetados para atuarem como um caminho otimizado aos dados espaciais com base em um conjunto definido de predicados sobre os atributos.

Ao longo do tempo, diversas pesquisas foram realizadas no intuito de melhorar as formas de indexação dos dados espaciais. As mais simples são as árvores binárias, como AVL, Red-Black e Splay Tree [Gaede and Günther 1998]. Após isto, diversas outras foram propostas, sendo as principais: KD-Tree [Bentley 1975], R-Tree [Guttman 1984], Hilbert R-Tree [Kamel and Faloutsos 1993], Grid [Nievergelt et al. 1984], e R+-Tree [Sellis et al. 1987].

O *SpatialHadoop*, que é a ferramenta a ser utilizada neste trabalho, utiliza os índices espaciais Grid, R-Tree e R+-Tree [Eldawy and Mokbel 2015b].

5. Trabalhos Relacionados

A comparação de desempenho na indexação de dados espaciais é tema recorrente na academia. Em 1993, [Ooi et al. 1993] apresentaram uma taxonomia sobre índices espaciais, entre eles o *Grid*, o *R-Tree* e o *R+-Tree*. Para os autores, independente do método utilizado, o desempenho da indexação é influenciada: pelo número de objetos espaciais por unidade de espaço; pelo tamanho dos objetos; e, pelo tamanho da base de dados.

Em [Teotônio 2008] é apresentada uma comparação do desempenho dos índices *R-Tree*, *Grid* e *Curvas de Hilbert* para consultas espaciais em bancos de dados geográficos relacionais, utilizando as ferramentas *PostgreSQL* com extensão espacial PostGIS, e *MySQL*. Os bancos de dados relacionais, também foram utilizados para comparação entre índices espaciais no trabalho apresentado por [Pant 2015].

Especificamente para *Big Geospatial Data*, [Eldawy et al. 2015] apresenta algumas técnicas de indexação através do *SpatialHadoop*. Segundo os autores, a tarefa de indexação no *SpatialHadoop* é proporcional ao tamanho da base.

Por fim, [Bachiega et al. 2017] apresentam um método focado na eficiência de custo para o processamento de *Big Geospatial Data* em ambiente de nuvem, levando em consideração o provisionamento do *cluster* baseado apenas no tamanho da base de dados a ser processada.

O presente trabalho, portanto, difere-se dos demais trabalhos já apresentados, porque propõe uma comparação de desempenho na indexação de *Big Geospatial Data*, através do *SpatialHadoop*, utilizando recursos oferecidos pela nuvem computacional. Além disso, este artigo busca indicar previamente a configuração mais adequada para otimizar o desempenho (tempo) da aplicação, baseando-se tanto nos tipos de dados contidos nos conjuntos de dados a serem utilizados, quanto nos parâmetros das consultas espaciais que serão realizadas.

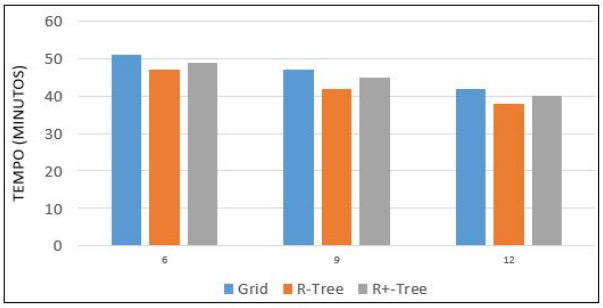


Figura 1. Tempo para Diferentes Tipos de Indexações.

6. Resultados

Um ambiente de testes foi configurado no provedor Microsoft Azure, utilizando o serviço *Azure HDInsight* [1], que é oferecido especificamente para a construção de aplicações que processam grandes volumes de dados. Foram utilizados três conjuntos de dados, todos extraídos do *OpenStreetMap* [2], conforme apresentado na Tabela 1.

Tabela 1. Conjuntos de Dados Utilizados nos Testes.

Conjunto de Dados	Conteúdo	Qtde. Registros
Pequena	Estradas mapeadas no mundo	20 milhões
Média	Construções mapeadas no mundo	115 milhões
Grande	Objetos mapeados no mundo	263 milhões

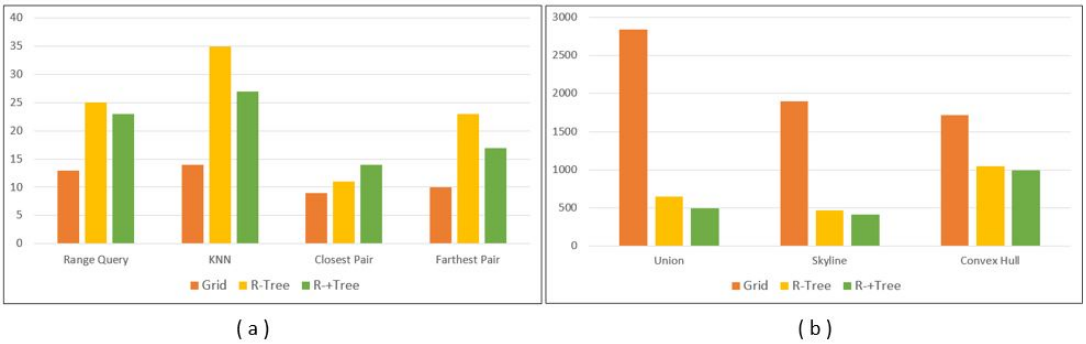
A Figura 1 apresenta o tempo (em minutos) para as indexações Grid, R-Tree e R+-Tree da Base Grande, variando a quantidade de nós do *cluster*. É possível notar que, o tempo reduzido não é proporcional a quantidade de nós adicionados. No caso da indexação R-Tree, por exemplo, embora a quantidade de nós tenha aumentado em 100%, passando de 6 nós para 12 nós, a redução de tempo foi apenas 20%.

Também foram executados testes, para todos os conjuntos de dados, com as seguintes tarefas executadas sequencialmente: 1- indexação *grid* do conjuntos de dados; 2- execução de consulta knn (*k-nearest neighbors*), com o valor de $k = 100$; e 3- execução da consulta por faixa (*range query*). Os tempos (em segundos) resultantes destas tarefas podem ser observados na Tabela 2. Embora a tarefa de indexação seja a mais onerosa, exigindo mais de 99% do tempo nos testes realizados, e também seja proporcional ao tamanho da base de dados, todas as tarefas seguintes são executadas em tempos similares, independente do tamanho da base.

Tabela 2. Tempos (em segundos) para Execução das Tarefas.

Tarefa	Base Pequena	Base Média	Base Grande
Indexar	602	3543	15361
KNN	8	10	10
Range	8	6	7

[1] www.microsoft.com/HDinsight
[2] http://spatialhadoop.cs.umn.edu/datasets.html



(a) (b)

Figura 2. Tempo para Diferentes Tipos de Indexações.

No entanto, a correta escolha da indexação, traz impactos significativos no desempenho. A Figura 2 apresenta o tempo (em segundos) para a execução de consultas e operações geográficas, após a Base Grande ter sido indexada. É possível notar que a indexação Grid tem um melhor desempenho para as consultas *Range Query*, *KNN*, *Closest Pair* e *Farthest Pair* (Figura 2a). Já para as operações *Union*, *Skyline* e *Convex Hull*, a indexação R+-Tree é a de melhor desemepenho (Figura 2b).

7. Conclusão

A indexação tem papel fundamental no desempenho de aplicações que processam *Big Geospatial Data*, uma vez que é a tarefa que exige maior poder computacional e tempo de processamento. Conforme demonstrado nos cenários de testes, a escolha correta da indexação é importante para a execução de maneira mais eficiente das operações e consultas geográficas.

O ambiente de computação em nuvem facilita o processamento de *Big Geospatial Data* uma vez que a demanda por recursos com alto poder computacional é obtida rapidamente. Nos testes realizados foi possível observar que o crescimento dos nós do *cluster* não reduz, de maneira proporcional, o tempo de processamento. Com isso, faz-se necessária ainda, a análise do custo para processamento deste tipo de aplicação em ambiente de nuvem.

A realização de testes com outras bases de dados, com outras indexações, e com a utilização de outras ferramentas que não só o *SpatialHadoop*, também são sugeridos como continuição deste trabalho. Desta forma, objetiva-se obter uma base de conhecimento suficiente para indicar a configuração mais performática, tanto em relação ao custo quanto em relação ao tempo, de acordo com os dados a serem processados e as consultas e as operações a serem realizadas.

Referências

Aji, A., Wang, F., Vo, H., Lee, R., Liu, Q., Zhang, X., and Saltz, J. (2013). Hadoop gis: a high performance spatial data warehousing system over mapreduce. *Proceedings of the VLDB Endowment*, 6(11):1009–1020.

Bachiega, J., Reis, M., Araujo, A., and Holanda, M. (2017). Cost optimization on public cloud provider for big geospatial data. *Proceedings of the 7th International Conference on Cloud Computing and Services Science*, pages 54–62.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517.

Eldawy, A., Alarabi, L., and Mokbel, M. F. (2015). Spatial partitioning techniques in spatialhadoop. *Proceedings of the VLDB Endowment*, 8(12):1602–1605.

Eldawy, A. and Mokbel, M. F. (2013). A demonstration of spatialhadoop: An efficient mapreduce framework for spatial data. *Proceedings of the VLDB Endowment*, 6(12):1230–1233.

Eldawy, A. and Mokbel, M. F. (2015a). The era of big spatial data. In *Data Engineering Workshops (ICDEW), 2015 31st IEEE International Conference on*, pages 42–49. IEEE.

Eldawy, A. and Mokbel, M. F. (2015b). Spatialhadoop: A mapreduce framework for spatial data. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*, pages 1352–1363. IEEE.

Gaede, V. and Günther, O. (1998). Multidimensional access methods. *ACM Computing Surveys (CSUR)*, 30(2):170–231.

Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. *SIGMOD international conference on Management of data*, 14(2).

Hoel, E. and Park, M. (2014). Big data: Using arcgis with apache hadoop. *Esri International Developer Summit*.

Kamel, I. and Faloutsos, C. (1993). Hilbert r-tree: An improved r-tree using fractals. *International Conference on Very Large Databases (VLDB)*.

Li, A., Yang, X., Kandula, S., and Zhang, M. (2010). Cloudcmp: comparing public cloud providers. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 1–14. ACM.

Nievergelt, J., Hinterberger, H., and Sevcik, K. C. (1984). The grid file: An adaptable, symmetric multikey file structure. *ACM Transactions on Database Systems (TODS)*, 9(1):38–71.

Ooi, B., Sacks-Davis, R., and Han, J. (1993). Indexing in spatial databases. *National University of Singapore*.

Pant, N. (2015). *Performance comparison of spatial indexing structures for different query types*. The University of Texas at Arlington.

Sellis, T., Roussopoulos, N., and Faloutsos, C. (1987). The r+-tree: A dynamic index for multi-dimensional objects. *International Conference on Very Large Databases (VLDB)*.

Teotônio, F. A. B. (2008). Comparacao do desempenho dosíndices r-tree, grades fixas, e curvas de hilbert para consultas espaciais em bancos de dados geograficos. *Dissertacao de Mestrado do Curso de Pos–Graduacao em Computacao Aplicada. Instituto Nacional de Pesquisas Espaciais-INPE, SP, Brazil*.

Yang, C., Yu, M., Hu, F., Jiang, Y., and Li, Y. (2017). Utilizing cloud computing to address big geospatial data challenges. *Computers, Environment and Urban Systems*, 61:120–128.