Uma Proposta de Comparação de Ferramentas para Análise de Grandes Conjuntos de Dados: Aplicação a Repositório sobre Monitoramento da Camada de Ozônio

Emilio Hoffmann de O.[1], **Andrea S. Charão** [1]

[1] Programa de Pós-Graduação em Informática
Universidade Federal de Santa Maria
Santa Maria, RS, Brasil

emiliohoffmann@gmail.com, andrea@inf.ufsm.br

Resumo. *Este artigo apresenta uma proposta de comparação de processamento de dados da camada de ozônio provenientes do Ozone Monitoring Instrument (OMI), utilizando o Apache Hadoop e outras ferramentas alternativas. A comparação deve verificar as diferenças, dificuldades e possibilidades oferecidas pelas ferramentas, para o problema em questão.*

1. Introdução

Atualmente, existem diversas ferramentas dispostas a auxiliar o processamento distribuídos de grandes conjuntos de dados. A plataforma Apache Hadoop [White 2015], amplamente utilizada, é desenvolvida em Java e trabalha sobre dados armazenados em um sistema de arquivos distribuído, chamado de *Hadoop Distributed File System* (HDFS). A plataforma implementa o modelo de programação MapReduce, que permite operações paralelas sobre o conjunto de dados distribuídos. Dadas as contínuas demandas da área, surgiram também ferramentas que podem ser utilizadas em conjunto ou paralelamente ao Hadoop, como Spark, Pig [Agneeswaran 2014].

Em ciências atmosféricas, há muitos repositórios de dados observacionais que são coletados há décadas e podem ser analisados com ferramentas modernas. No trabalho de [Steffenel et al. 2016], utiliza-se um repositório sobre monitoramento da camada de ozônio terrestre, para identificar os *Ozone Secondary Effects* (OSE) que são caracterizados pela redução da camada de ozônio. A identificação desses fenômenos é importante para que seja possível alertar a população de uma região sobre o aumento da radiação ultravioleta.

O trabalho citado emprega um *middleware* de computação pervasiva e distribuída, que é vantajoso nos casos em que não se dispõe de uma infraestrutura computacional dedicada [Steffenel et al. 2016]. No presente trabalho, tem-se por objetivo explorar outras alternativas voltadas para ambientes dedicados, buscando compará-las no contexto do problema em questão .

2. Dados de Monitoramento da Camada de Ozônio

Os dados utilizados na análise serão do Ozone Monitoring Instrument (OMI), que usa sensores a bordo de satélites. Esses dados são referentes a leituras da camada de ozônio da terra e são disponibilizados pela *National Aeronautics and Space Administration* (NASA) em [NASA 2016].

Esses dados são gerados diariamente, gerando cada dia um arquivo com aproximadamente 2700 linhas de dados, chegando próximo a 1 milhão de linhas durante o ano, contendo informações da espessura da camada de ozônio de acordo com as coordenadas informadas. O grande volume de dados torna o processamento demorado e complicado, de forma que este problema poderia ser beneficiado pela utilização de ferramentas para processamento distribuído de dados.

3. Metodologia

No trabalho a ser desenvolvido, primeiramente serão desenvolvidas aplicações para a análise dos dados citados na seção 2, usando apenas o Apache Hadoop. Serão realizadas diversas operações nos dados, como: média, desvio padrão, variância, entre outras. Outras ferramentas devem ser estudadas para a realização do artigo, entre elas estão Spark e Pig.

Em [Gu e Li 2013], os autores definem Spark como um *framework* para *clusters* que utiliza o modelo de programação MapReduce, foi desenvolvido para melhorar as operações iterativas do Hadoop. Através da sua estrutura de dados o Spark permite que os dados e resultados intermediários sejam armazenados em cache e reutilizados durante todo o processo iterativo.

Outra ferramenta a ser estudada é a Pig, descrita em [Olston et al. 2008] como sendo uma linguagem desenvolvida entre linguagens declarativas como SQL e linguagens procedurais de baixo nível no estilo de MapReduce e foi desenvolvida para rodar acima do Apache Hadoop.

Após a primeira análise com o Apache Hadoop, serão desenvolvidas as mesmas aplicações utilizando as demais ferramentas a serem estudadas. Após o desenvolvimento será realizada a comparação, analisando as dificuldades encontradas em cada ferramenta, as facilidades, etc. As métricas consideradas deverão compreender aspectos de desempenho (por exemplo, tempo de execução) e de engenharia de software (por exemplo, número de linhas de código).

Referências

Agneeswaran, V. (2014). *Big data analytics beyond Hadoop*. Upper Saddle River, NJ: Pearson Education.

Gu, L. e Li, H. (2013). Memory or time: Performance evaluation for iterative operation on hadoop and spark. In *High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC), 2013 IEEE 10th International Conference on*, pages 721–727. IEEE.

NASA (2016). Ozone: Multi-mission ozone measurements. https://ozoneaq.gsfc.nasa.gov/data/ozone/.

Olston, C., Reed, B., Srivastava, U., Kumar, R., e Tomkins, A. (2008). Pig latin: a not-so-foreign language for data processing. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1099–1110. ACM.

Steffenel, L. A., Pinheiro, M. K., Pinheiro, D. K., e Perez, L. V. (2016). Using a pervasive computing environment to identify secondary effects of the antarctic ozone hole. *Procedia Computer Science*, 83:1007–1012.

White, T. (2015). *Hadoop: The Definitive Guide*. O'Reilly Media Inc.