Estudo das tecnologias aplicadas no conceito de BigData

Maurício dos Santos Dessuy[1], **Edson L. Padoin**[1]

[1]Universidade Reg. do Noroeste do Estado do Rio G. do Sul (UNIJUI)
Ijuí - RS - Brasil

mauriciodessuy@hotmail.com, padoin@unijui.edu.br

Resumo. *O processamento de grandes volumes de dados (Big Data) e seu armazenamento distribuído vem aumentando gradualmente o uso da rede. Nesse contexto, diferentes empresas e instituições passaram a estudar formas de como armazenar e processar estes grandes volume de dados, obtendo valor e conhecimento a partir deles. O objetivo deste trabalho é realizar um estudo introdutório sobre as tecnologias e ferramentas que auxiliam desenvolvedores na solução de problemas e uso de Big Data.*

1. Introdução

A quantidade de dados que os sistemas e usuários geram vem crescendo de forma exponencial. Em 2012 foram gerados cerca de $2,5$ Exabytes de dados por dia. Este volume de dados é tão elevado, fazendo com que a quantidade de dados que trafegam na internet atualmente é maior do que se tinha armazenado no mundo inteiro cerca de 20 anos atrás [McAfee et al. 2012]. Nesse sentido, um desafio dos atuais sistemas de informações é como armazenar e processar estes volumes de dados.

Este grande conjunto de dados é atualmente denominado de BIGDATA, podendo armazendar e processar dados gerados a partir de diversas fontes, como computadores, celulares, câmeras, sensores e diversos outros dispositivos que acessam a internet [Deighton 2019]. Partindo desta premissa, este artigo pretende promover um embasamento teórico acerca de BIGDATA, bem como comparar as ferramentas e soluções utilizados atualmente. Assim, inicialmente apresentamos uma conceituação e um comparativo das tecnologias disponíveis para BIGDATA. Na sequência, na Sessão 2 são apresentados os conceitos e desafios envolvidos neste contexto, bem como o framework Hadoop. Na Sessão 3 são destacadas algumas tecnologias e ferramentas de BIGDATA. A Sessão 4 destacam-se os trabalhos relacionados. O trabalho é finalizando na Sessão 5 com algumas propostas de trabalhos futuros.

2. Big Data

BIGDATA pode ser definido como uma nova geração de tecnologias e arquiteturas pensadas a partir do conceito $5V$. Foram projetadas para extrair *valor* de grandes *volumes* de dados, que possuem uma *variedade* de fontes, uma alta *velocidade* de geração além da sua *veracidade* [Gandomi and Haider 2015]. Atualmente, existem amplas discussões sobre quando uma quantidade de dados começa a ser considerada BIGDATA. No entanto, ainda não existe um amplo consenso sobre estas medidas. Por outro lado, para processar estes grandes volumes de dados necessita-se de abordagens diferentes para *processamento* e *armazenamento*, uma vez que os sistemas de bancos de dados tradicionais, denominados de relacionais, não apresentam boa eficiência quando utilizados neste novo ambiente.

Por conta disto, diferentes frameworks começam ganhar mercado almejando auxiliar nas tarefa de *processamento* e *armazenamento* de BIGDATA, dentre os quais destaca-se o Apache Hadoop.

O Hadoop é um framework desenvolvido na linguagem Java para *processamento* e *armazenamento* de dados de forma distribuída. Sua principal premissa é a escalabilidade horizontal, que, em caso de aumento da demanda, prevê um aumento da quantidade de servidores ou nodos conectados no cluster, diferentemente das metodologias tradicionais que preveem um escalabilidade vertical, onde o poder de processamento dos servidores em questão é incrementado [Lam 2010].

O framework Hadoop foi conceitualmente planejado para utilizar o sistema de arquivos distribuídos Hadoop Distributed File System (**HDFS**) para o *armazenamento* de dados e o framework **MapReduce** para o *processamento* de dados.

O HDFS é um sistema de arquivos tolerante a falhas projetado para executar em hardware de baixo custo. Sua estrutura é organizada em forma de clusters, sendo estes divididos em dois grupos de nós. Os nós *NameNodes* que possuem a função de gerenciar o sistema de arquivos e nós *DataNodes* que possuem a função de armazenar os dados salvos [White 2012]. Assim, almejando garantir a integridade e disponibilidade dos dados, o HDFS implementa um nível de replicação que define a quantidade de vezes que os mesmos são replicados entre os nós do cluster. Deste modo, caso um nó apresente falhas, o HDFS consegue acessar os dados de outro nodo ativo e continua o seu funcionamento.

Por outro lado, para o processamento dos dados em BIGDATA, o Hadoop utiliza a tecnologia de MapReduce. Este framework tem como função fornecer um modelo de programação para tratamento dos dados. As tarefas são executadas baseadas nas funções de sumarizar e reduzir os dados afim de obter um resultado sobre os dados processados [Chu et al. 2007].

Assim, todo o processo do MapReduce é dividido em duas fases. Na primeira fase - Map - os dados são mapeados, tratados e agrupados por uma chave que permita a indexação das informações. Nesta fase, formam-se tuplas contendo uma chave e um conjunto de valores. Na segunda fase - Reduce - é realizado o processamento nos dados agrupados a fim de torná-los úteis, como por exemplo, realizar uma soma dos valores das chaves mapeadas [Venner 2009].

3. Tecnologias de BIGDATA

Diferentes ferramentas e frameworks tem sido desenvolvido para *processamento* e *armazenamento* de dados no contexto de BIGDATA. Dentre as quais destacam-se:

3.1. Tecnologias para Processamento

- **Hadoop Yarn** - Foi implementado a partir da versão 2.0 do Hadoop. Esta ferramenta tem como função separar as funcionalidades de gerenciamento de recursos e agendamento de tarefas dentro do framework [Kulkarni and Khandewal 2014];
- **Hadoop Pig** - É uma linguagem de alto nível para análise de grandes quantidades de dados [Lam 2010]. É uma ferramenta empregada para auxiliar desenvolvedores na solução de problemas envolvendo BIGDATA; e

- **Hadoop Spark** - É uma interface de programação para tratamento e processamento de dados. A ferramenta Spark tem sido bastante utilizada dado o seu desempenho na execução de aplicações, alcançando eficiências de até 40 vezes maiores que o Hadoop, isto devido a sua característica de manter os dados armazenados em memória [Zaharia et al. 2010].

3.2. Tecnologias para Armazenamento

- **HBase** - É um framework para Hadoop que disponibiliza diversas ferramentas para auxiliar no armazenamento dos dados. Dentre elas destaca-se a função de proporcionar um data warehouse para que as aplicações possam armazenar seus dados de forma não estruturada (NoSql) [Lam 2010].
- **Hive** - É uma ferramenta Hadoop que permite um armazenamento de dados no formato estruturado. Com ela é possível a criação de tabelas relacionais através da linguagem Hive Query, esta que se assemelha bastante a linguagem SQL [Lam 2010].
- **MongoDB** - É um banco de dados não relacional (NoSQL) orientado a documentos. Sua principal diferença com os bancos de dados relacionais é que nele não existe o conceito de linhas e tabelas, assim os dados são inseridos no banco em forma de documentos. assim como não existe relacionamentos pois as informações referentes a um objeto são salvas em apenas um documento.

4. Trabalhos Relacionados

No artigo de [Qureshi and Koubaa 2019] é apresentado um estudo sobre a eficiência energética no uso do Hadoop em computadores de placa única (SBC), onde foi verificado que as arquiteturas de computadores SBC's baseadas em ARM apresentam um baixo custo no gasto de energia, tornando-os assim altamente eficientes na utilização de recursos computacionais.

Em [Dadheech et al. 2018] os autores desenvolveram uma aplicação de otimização de consultas para reduzir o tempo na busca de dados no ambiente de BIG-DATA. Após o desenvolvimento os autores realizaram testes de desempenho da nova metodologia com as técnicas de agendamento FIFO, HFS e HCS para verificar se a nova tecnologia apresenta um melhor desempenho.

Com a premissa de melhorar a performance no gerenciamento de recursos do Hadoop Yarn os autores do artigo [Yao et al. 2019] desenvolveram dois novos algoritmos de agendamento para o Hadoop Yarn, sendo eles o HaSTE e HaSTE-A, onde ambos os algoritmos tiveram uma melhora no tempo total de execução de processos MapReduce.

No estudo de [Óskarsdóttir et al. 2019] os autores demonstram como incluir a grande quantidade de dados provenientes de chamadas telefônicas e dados de redes sociais no ranqueamento de crédito. Ao fim do estudo é apontado que o uso destes recursos pode amplificar o mercado de empréstimos, pois a falta de conhecimento sobre o perfil de consumo das pessoas gera um alto custo para iniciantes no mercado de crédito.

5. Considerações Finais e Trabalhos Futuros

Tendo analisado os frameworks e as ferramentas disponíveis, percebe-se que a larga utilização do framework Apache Hadoop sendo que o emprego ela é fundamental para

geração de valor a partir de BIGDATA. Os sistemas em execução precisam atuar em uma enorme quantidade de dados fazendo com que novas metodologias de armazenamento e processamento sejam necessárias.

Um tecnologia que tem apresentado aumento em sua utilização é o Hadoop. Isto também, devido ao número de ferramentas que facilitam a resolução de problemas envolvendo neste novo contexto de dados. Tais ferramentas diferem das ferramentas tradicionais tanto em armazenamento quanto em processamento.

Como futuros trabalhos, pretende-se realizar a instalação destas ferramentas e realizar uma comparação de desempenho e eficiência tanto no armazenamento quanto no processamento de dados não estruturados. Nesse contexto, pretende-se verificar qual das tecnologias apresenta melhor desempenho em diferentes configurações de BIGDATA.

Referências

Chu, C.-T., Kim, S. K., Lin, Y.-A., Yu, Y., Bradski, G., Olukotun, K., and Ng, A. Y. (2007). Map-reduce for machine learning on multicore. In *Advances in neural information processing systems*, pages 281–288.

Dadheech, P., Goyal, D., Srivastava, S., and Kumar, A. (2018). Performance improvement of heterogeneous hadoop clusters using query optimization.

Deighton, J. (2019). Big data. *Consumption Markets & Culture*, 22(1):68–73.

Gandomi, A. and Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144.

Kulkarni, A. P. and Khandewal, M. (2014). Survey on hadoop and introduction to yarn. *International Journal of Emerging Technology and Advanced Engineering*, 4(5):82–87.

Lam, C. (2010). *Hadoop in action*. Manning Publications Co.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., and Barton, D. (2012). Big data: the management revolution. *Harvard business review*, 90(10):60–68.

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., and Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing*, 74:26–39.

Qureshi, B. and Koubaa, A. (2019). On energy efficiency and performance evaluation of single board computer based clusters: A hadoop case study. *Electronics*, 8(2):182.

Venner, J. (2009). *Pro hadoop*. Apress.

White, T. (2012). *Hadoop: The definitive guide*. "O'Reilly Media, Inc.".

Yao, Y., Gao, H., Wang, J., Sheng, B., and Mi, N. (2019). New scheduling algorithms for improving performance and resource utilization in hadoop yarn clusters. *IEEE Transactions on Cloud Computing*.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., and Stoica, I. (2010). Spark: Cluster computing with working sets. *HotCloud*, 10(10-10):95.