CONIC SEMESP

15º Congresso Nacional de Iniciação Científica

TÍTULO: PLATAFORMA DE BIG DATA COM ÊNFASE EM BAIXO CUSTO

CATEGORIA: EM ANDAMENTO

ÁREA: CIÊNCIAS EXATAS E DA TERRA

SUBÁREA: COMPUTAÇÃO E INFORMÁTICA

INSTITUIÇÃO: FACULDADE DE TECNOLOGIA DE SÃO JOSÉ DO RIO PRETO

AUTOR(ES): HANNA CAROLINA BARBOSA, ARIAN AUGUSTO BOTINE, FLÁVIO MARTINHO DA SILVA

ORIENTADOR(ES): JOSÉ APARECIDO DE AGUIAR VIANA

COLABORADOR(ES): SÉRGIO RICARDO BORGES JÚNIOR

Realização: SEMESP

Apoio: UNAERP Universidade de Ribeirão Preto Campus Ribeirão Preto - Campus Guarujá

RESUMO

Empresas e grupos de pesquisas estão cada vez mais explorando as possibilidades oferecidas na área de *big data*, porém podem não disponibilizar da arquitetura necessária e investimento suficiente para a construção de um *cluster*. Uma alternativa é utilizar máquinas com *hardware* comum para construção da plataforma de *big data* com *softwares open-source,* tais como o sistema operacional *Linux* e uma distribuição do *framework Hadoop*. Nesta perspectiva, o presente projeto objetiva propor uma abordagem de *big data* com *hardware* e *software* de baixo custo para viabilizar pesquisas acadêmicas ou aplicações comerciais que envolvam grandes volumes de dados necessitando de alto desempenho de processamento.

INTRODUÇÃO

O *Hadoop* é uma ferramenta amplamente utilizada por grandes corporações para aplicações de *big data,* destacando-se sua capacidade para armazenamento e processamento de grande volume de dados, além de oferecer confiabilidade e elevado grau de tolerância a falhas [1]. Dessa forma, o *framework* é capaz de unir de forma eficiente grande capacidade de armazenamento e processamento de informações para obter um bom desempenho em análises específicas e sendo possível sua aplicação em *hardware* de baixo custo com base em resultados já apresentados por sua idealizadora, a *Apache Foundation* [2]. Sendo assim, o *Hadoop* pode ser aplicado na arquitetura de um *cluster* construído com máquinas de *hardware* comum, podendo explorar as funcionalidades proporcionadas pela ferramenta, as quais se baseiam em processamento distribuído. A grande vantagem está justamente na possibilidade de reduzir significativamente os custos da implementação de um *cluster*, que muitas vezes acaba sendo necessário um grande investimento [3]. Para isso, o presente projeto utilizará uma plataforma para a aplicação do *Hadoop* em um *cluster* real. No mercado é possível encontrar algumas implementações do *framework*, sendo que uma das mais comuns é a plataforma *Hortonworks*. Essa plataforma proporciona de uma maneira simplista e completa a aplicação do *Hadoop* juntamente com um conjunto de ferramentas que além de serem capazes de processar e gerenciar grandes quantidades de dados paralelamente também garantem segurança, disponibilidade, confiabilidade entre outras características para obtenção de resultados satisfatórios atendendo a necessidades complexas [4].

OBJETIVOS

O objetivo do projeto é propor uma abordagem de *big data* com *hardware* e *software* de baixo custo para viabilizar aplicações/pesquisas acadêmicas e comerciais que envolvam grandes volumes de dados e que necessitam de alto desempenho de processamento. Dessa forma, busca-se a utilização de máquinas com *hardware* comum, sistema operacional e ferramentas de código aberto para prover as características mencionadas.

METODOLOGIA

O projeto envolve estudo de campo sobre o *Hadoop* utilizando a plataforma *Hortonworks* que traz também outros recursos para aplicação em *big data*. O levantamento bibliográfico relacionado ao *framework* e a própria plataforma possibilitou a construção de um modelo virtualizado através do *VirtualBox* [5], para verificar a viabilidade do projeto. Após essa validação foi realizado a construção de um modelo físico da solução proposta, voltada ao baixo custo, para efetuar testes de desempenho e adquirir as métricas responsáveis pela análise dos resultados parciais para a conclusão do projeto. Na construção do cluster foi utilizado máquinas doadas pela comunidade e faculdade que possuem *hardware* com processador Intel Pentium 4 e memória principal de 2 GBytes, conectadas pela rede através de um *switch*.

DESENVOLVIMENTO

Foram realizadas pesquisas sobre a montagem de um cluster e utilização do *Hadoop* com foco em baixo custo. Depois de vários testes definiu-se a distribuição da empresa *Hortonworks* como plataforma de *big data* pelos recursos que disponibiliza juntamente com o *Hadoop*, por se tratar de uma plataforma aberta e simplista. Com isso, procedeu-se a montagem da arquitetura, organização do *cluster* físico e configuração da plataforma utilizando máquinas com *hardware* de baixo custo. Estudos foram feitos para a preparação e configurações de testes utilizando *MapReduce*, *framework* para processamento paralelo, por enquanto os testes que foram realizados aplicaram o algoritmo de contagem de palavras em um arquivo de 1 GByte que após seu processamento gerou-se um relatório com o tempo total do processamento e de cada fase do *MapReduce*. As fórmulas (1) e (2) foram utilizadas como métricas para avaliar o desempenho do *Hadoop*:

$$S(p) = T(1)/T(p) \quad (1)$$

$$E(p) = S(p)/p \quad (2)$$

Sendo: *S(p) speedup*; *E(p)* eficiência; *T(1)* o tempo em um nó; *T(p)* o tempo de execução em *p* nós.

RESULTADOS PRELIMINARES

Os resultados foram obtidos utilizando a média de 20 testes consecutivos com o mesmo padrão para cada nó adicionado. A Figura 1 retrata os resultados baseando-se nas métricas explicadas anteriormente.

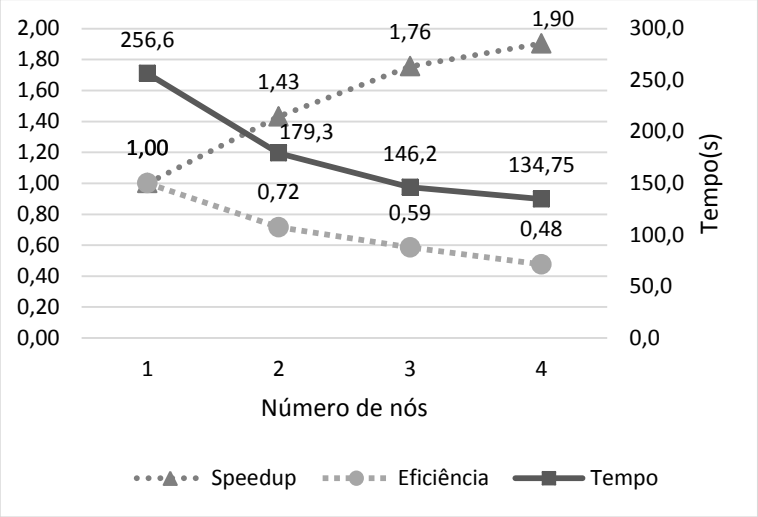

Figura 1 - Demonstração de tempo (s), Speedup e Eficiência por quantidade de máquinas no cluster

Pode-se perceber uma diminuição de tempo no total de 47,48% entre a situação de 1 nó e a de 4 nós. Apesar de ser possível obter resultados comparativos com o uso das métricas explicadas anteriormente ainda é necessário realizar estudos e testes sobre o sistema de arquivos distribuído do *Hadoop* e as outras ferramentas que englobam a plataforma para realçar a funcionalidade da arquitetura em si.

FONTES CONSULTADAS

[1] WHITE, Tom. Hadoop: The Definitive Guide. Second Edition. 2011. Editora: O'Reilly Media, Sebastopol – CA

[2] Apache Hadoop Project (2015, Mar), Available: http://hadoop.apache.org

[3] F. Martinho et al., SICT Virtualização de Cluster HPC. Página 139 (2014)

[4] Hortonworks Documentation (2015, Abr), Available: http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.2.6/index.html

[5] Virtualbox Documentation (2015, Jun), Available: https://www.virtualbox.org/wiki/Documentation

[6] M. Kontagora e H. Gonzalez. IEEE Benchmarking a MapReduce Environment on a Full Virtualisation Platform, 1-6 (2010)