Análise da plataforma HDFS como serviço de armazenamento de alta disponibilidade

Guilherme de Oliveira Ferreira[1]; Lucas Austro Fernandes Bazante Junior[1]; Wagner Neiva Costa[1]; Professor Me. Francisco Supino Marcondes[1];

[1]Instituto Federal de Educação Ciência e Tecnologia de São Paulo – *Campus* São Paulo;

Objetivo

O objetivo desta pesquisa visa analisar o potencial estratégico na adoção do HDFS – sistema distribuído de arquivos – para suportar o gerenciamento de dados com alta disponibilidade no âmbito da Tecnologia da Informação e Comunicação (TIC).

Introdução

A revolução no conceito de Big Data passa pela exploração de temas que gerenciem grandes volumes de dados e, dentre estes, sistemas de arquivos distribuídos, tal como o *Hadoop Distributed File System* (HDFS) – estrutura baseada no armazenamento e gerenciamento de dados em um sistema distribuído de arquivos –, tem sido abordados como recorrente tema de estudos (1,2,3). O HDFS – projetado pela Apache (3) (Figura 1) e baseado no trabalho anterior no *Google Distributed Files System* – tem como principais objetivos: uma rápida detecção e recuperação de falhas, acesso a dados de fluxo (*MapReduce*), modelo de simultaneidade simples e robusto, escalabilidade para armazenar e processar grandes quantidades de dados, economia pela distribuição e processamento de dados entre cluster (4). A atual implementação do HDFS dá suporte a alta disponibilidade de metadados através de um protocolo de replicação eventualmente consistente (5), este, realiza-se separado em grandes blocos e distribuído em clusters de várias máquinas para evitar erros através das replicações dos arquivos e garantindo dados sempre disponíveis (6). A Figura 1 ao lado descreve a arquitetura do HDFS.

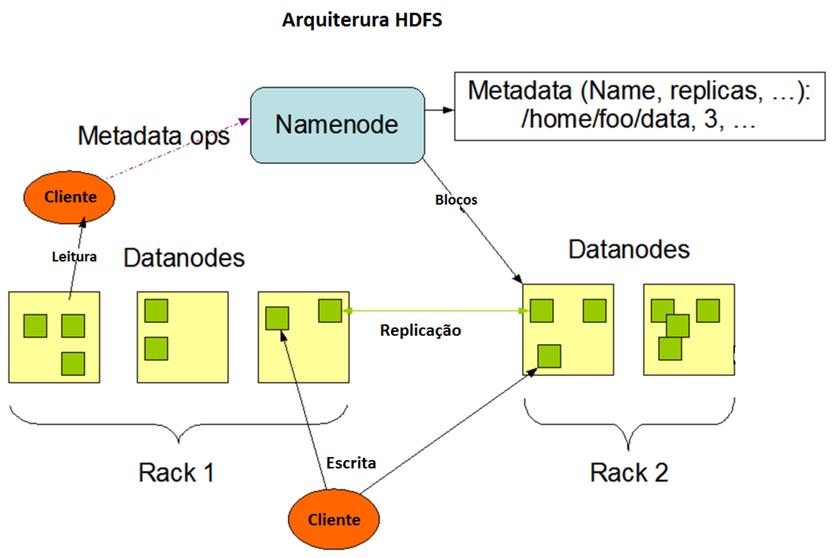
Materiais e Métodos

Inicialmente, a base metodológica deste trabalho se dará por dois tipos de técnicas de pesquisa: bibliográfica – que permite "ao investigador a cobertura de uma gama de fenômenos muito mais ampla do que aquela que poderia pesquisar diretamente" e documental, pois "os documentos constituem fonte rica e estável de dados" (7). Espera-se, também, uma abordagem qualitativa para descrever experiências observadas para a apresentação de uma conclusão assertiva acerca do tema.

Resultados preliminares

O presente trabalho pretende apurar a aplicação do HDFS como um diferencial competitivo perante resultados de demais soluções que apresentem benefícios em relação a armazenamento com alta disponibilidade. Ademais, neste, estudo foi possível obter os resultados (assumindo 10.000 nós capazes de armazenar 1TB cada): **Nível aceitável de perda de dados** como 1 hora – todos os dados criados ou atualizados em DFS há uma hora atrás ou antes são garantidos para serem recuperados em caso de falhas do sistema; **Nível de inatividade aceitável** como 2 horas – a falha DFS requer recuperação manual do sistema. Neste último, o sistema está garantido a estar disponível novamente, no mais tardar, 2 horas após o início da recuperação. (8)

Figura 1: Arquitetura HDFS



Conclusão

O presente trabalho pretende possibilitar uma análise mais aprofundada e assertiva da observância do HDFS permitir que aplicações trabalhem com milhares de nós em cluster se mostra mais eficaz dentre as soluções de serviços de armazenamento com alta disponibilidade. Assim, podendo refletir os potenciais aspectos estratégicos para a sua adoção no âmbito da Tecnologia da Informação e Comunicação (TIC).

Referências

(1) HANSON, J. Uma introdução ao Hadoop Distributed Files System. **IBM developerWorks Brasil**, Brasília, ago., 2013. Disponível em: <https://www.ibm.com/developerworks/br/library/wa-introhdfs/index.html>. Acesso em: 28 set. 2017.

(2) SHVACHKO, K.; KUANG, H.; RADIA, S.; CHANSLER, R. The Hadoop Distributed File System. In: Proceedings of MSST2010, 26., 2010, Nevada. **Anais**… Nevada: IEEE Press, 2010. p. 1-10.

(3) APACHE Software Foundation. **HDFS High Availability**, 2017. Disponível em: < https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HDFSHighAvailabilityWithNFS.html> Acesso em: 20 set. 2017.

(4) Borthakur, D. **HDFS Architecture**: HDFS Architecture Guide, 2013. Disponível em: <https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html>. Acesso em: 20 set. 2017.

(5) Mohammed A. F; Humbe V. T.; Chowhan S. S. A review of Big Data Environment and its Related Technologies. In: International Conference on Information Communication and Embedded Systems (ICICES), 5., 2016, Tamilnadu. **Anais**… Nevada: IEEE Press, 2016.

(6) Hakimzadeh, K.; Sajjad, H. P.; Dowling, J. Scaling HDFS with a Strongly Consistent Relational Model for Metadata. In: Proceedings of the 14th IFIP International Conference on Distributed Applications and Interoperable Systems, 14., 2014, Berlin. **Anais…** New York: Springer-Verlag New York, 2014. p. 38-51.

(7) GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

(8) APACHE Software Foundation. **The Hadoop Distributed File System requirements**, 2011. Disponível em <https://wiki.apache.org/hadoop/DFS_requirements>. Acesso em: 20 set. 2017.