UNICEPLAC

Centro Universitário do Planalto Central Apparecido dos Santos

Curso de Sistemas de Informação

Trabalho de Conclusão de Curso

Sistema de arquivos distribuídos HDFS

Hadoop File System

Brasília-DF

2019

UNICEPLAC

André Rodrigues de Melo Júnior

João Paulo da Silva Pereira

Wilton Bezerra da Silva

Sistema de arquivos distribuídos HDFS

Hadoop File System

Artigo apresentado como requisito para conclusão do curso de Bacharelado em Sistemas de Informaçãopelo Centro Universitário do Planalto Central Apparecido dos Santos – Uniceplac.

Orientador: Prof.Ms.Ararigleno Almeida Fernandes

Brasília-DF

2019

UNICEPLAC

André Rodrigues de Melo Júnior

João Paulo da Silva Pereira

Wilton Bezerra da Silva

Sistema de arquivos distribuídos HDFS

Hadoop File System

Artigo apresentado como requisito para conclusão do curso de Bacharelado em Sistemas de Informação pelo Centro Universitário do Planalto Central Apparecido dos Santos – Uniceplac.

Gama, 22 de Novembro de 2019.

Banca Examinadora

Prof. Ms. Arariglenо Almeida Fernandes
Orientador

Prof. Ms. Osman Oliveira
Examinador

Prof. Ms. Jorge Santos
Examinador

UNICEPLAC

Sistema de arquivos distribuídos HDFS

Hadoop File System

André Rodrigues de Melo Júnior[1]

João Paulo da Silva Pereira[2]

Wilton Bezerra da Silva[3]

Resumo:
Na era digital as novas tecnologias geram constantemente elevadas quantidades de dados, podendo esses serem advindos de diversa fontes, formas e formatos, surge nesse sentido o conceito de *Big Data*. Esse grande volume de dados usado da maneira correta, pode ser de grande valor, gerando informações e insights importantes para o negócio ou organização. O uso de ferramentas como o Hadoop se faz necessário para processar, gerir e transformar esses dados em informação. Este artigo apresenta alguns aspectos relacionados a utilização da ferramenta Hadoop e seus complementos para a entrega de análises de dados.

Palavras-chave: Dados. *Big Data*. Ferramentas. Hadoop.

Abstract:
In the digital age the new technologies generate constantly high amounts of data, which can come from different sources, forms and formats, the concept of *Big Data* arises in this sense. This large volume of data used in the correct way, can be of great value, generating information and insights important to the business or organization. The use of tools such as Hadoop is necessary to process, manage and transform this data into information. This article presents some aspects related to the use of the Hadoop tool and its complements for the delivery of data analysis.

Keywords: Data. *Big Data*. Tools. Hadoop.

[1]Graduando do Curso Sistemas de informação, do Centro Universitário do Planalto Central Apparecido dos Santos–Uniceplac. E-mail: andre.rodrigues27@outlook.com.
[2]Graduando do Curso Sistemas de informação, do Centro Universitário do Planalto Central Apparecido dos Santos–Uniceplac. E-mail: dasilvapereirajoaopaulo@gmail.com.
[3]Graduando do Curso Sistemas de informação, do Centro Universitário do Planalto Central Apparecido dos Santos–Uniceplac. E-mail: wilton.bezerra.silva@gmail.com.

UNICEPLAC

1. INTRODUÇÃO

Com o advento de novas tecnologias e a evolução constante dos meios de comunicação pode-se observar que praticamente todos eles estão gerando constantemente grandes volumes de dados, porém os dados em si não agregam valor enquanto separados, mas sim por meio do conjunto deles é possível obter informações que de algum modo serão usadas em favor de uma organização ou negócio. Atualmente o volume de dados cresce a cada segundo nas inúmeras plataformas e ambientes online, e de acordo com a revista Forbes[4] até 2025 existirão 180 zetabytes de dados no mundo.

Nesse contexto, Big Data refere-se ao agrupamento de dados (dataset) cujo tamanho está além da habilidade das ferramentas típicas de Sistemas Gerenciador de Banco de Dados capturar, gerenciar e analisar. O conceito é intencionalmente subjetivo e incorpora uma definição genérica de como quão grande um conjunto de dados é necessário a fim de ser considerado Big Data. Os dados, surgem de todos os lados, sendo estes oriundos de diversos lugares desde sensores e câmeras de segurança, que praticamente cobrem todas as cidades, até as fotos clicadas por bilhares de smartphones, posts de textos e dos chats nas redes sociais, sem falarmos de todos os computadores existentes e conectados na internet. Segundo Santanchè (2014), o *Big Data*, embora seja tratado por muitos como solução, em si é um problema, pela quantidade e diversidade de dados, o que reivindica o uso de ferramentas específicas de análises de *Big Data*.

Segundo Taurion em seu livro *Big Data* (2013) a complexidade do *Big Data* não está apenas no tratamento desse volume de dados massivos, que vem de variadas fontes e que demandam alta velocidade de processamento, mas na revisão e criação de processos na busca por um valor. As ferramentas também conhecidas como *Big Data Analytics* permitem a gestão de grandes quantidades de dados além da manipulação de vários tipos de dados advindos de diversas fontes com diversificados formatos, algo que os sistemas tradicionais de banco de dados são limitados para resolver.

Pensando nisso, o presente artigo irá apresentar o Hadoop, uma ferramenta *open source* mantida pela *Apache Software Foundation*, projetada para garantir larga escalabilidade, tolerância à falhas, confiabilidade e velocidade na análise de grande volume de dados partindo de um único servidor a até um cluster com milhares de máquinas usando computação

[4]Gil Press. Forbes 2019. 6 previsões para o mercado de US$203 bilhões em *Big Data* Analytics. Disponível em: https://www.forbes.com/sites/gilpress/2017/01/20/6-predictions-for-the-203-billion-big-data-analytics-market/

distribuída. Nesse sentido o trabalho tem por objetivo a construção de um cluster físico Hadoop e a demonstração de cada passo da computação distribuída para análises de grandes volumes de dados com o uso do sistema de *data warehouse* Apache Hive que foi desenvolvido para compor o ecossistema Hadoop, além do uso de ferramentas de *Business Intellingence* para análise final dos dados.

Para chegar ao resultado final dessa pesquisa, foram usados conteúdos de diversas fontes nos diversos meios de pesquisa como internet, livros e artigos já publicados além da orientação de especialistas no assunto.

2. *BIG DATA* E HADOOP

Esta seção apresentará em seus tópicos os conceitos de *Big Data* e suas aplicações, bem como a necessidade do uso de ferramentas como o Hadoop e seus complementos para melhor desempenho em análises de dados.

2.1. Conhecendo o *Big Data*

A competição global exige sempre que as empresas busquem inovações e aperfeiçoamentos de seus produtos e processos, para isso é necessário sempre verificar se as ações aplicadas estão gerando valor ao negócio. O mundo se encontra na era da informação em que as análises de dados em uma empresa ou organização podem ser aplicadas em praticamente todas funções, processos de negócios, decisões e ações. Usando uma abordagem correta as empresas podem usar a análise de dados para impulsionar seu negócio e até redefinir a experiência do cliente.

Com as constantes evoluções tecnológica o volume de dados gerados vem aumentando a cada segundo e de forma exponencial. As análises de dados antes feitas em âmbito empresarial tomam agora maiores proporções sendo estas advindas de diversas fontes e em diversos formatos, o desafio agora é saber como usar e tratar esse grande volume de dados e a partir desse cenário surge o *Big Data.*

O termo *Big Data* é referente ao processamento de conjunto de dados extremamente grandes, que por sua variedade e tamanho não podem ser processados utilizando as ferramentas mais comuns de análise de dados. A empresa de consultoria Gartner[5] por sua vez define o *Big Data* como,

> o termo adotado pelo mercado para descrever problemas no gerenciamento e

[5]Gartner 2019. *Big Data*. Disponível em: https://www.gartner.com/it-glossary/big-data/

> processamento de informações extremas as quais excedem a capacidade das tecnologias de informações tradicionais ao longo de uma ou várias dimensões (Gartner apud Silva, 2013).

O termo *Big Data* não pode ser definido apenas como um grande volume de dados, segundo INTEL (2013) por se tratarem de dados extremamente amplos, necessitam de ferramentas específicas para lidar com esses grandes volumes de forma que façam análises que gerem informações úteis e em tempo hábil.

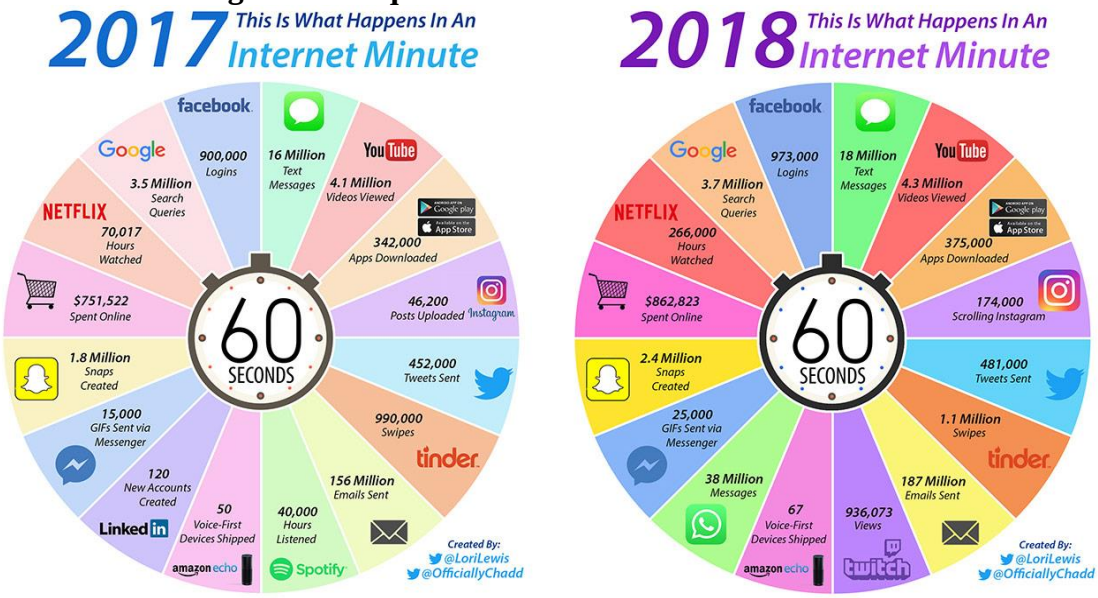
> "O valor real do *Big Data* está no insight que ele produz quando analisado – buscando padrões, derivando significado, tomando decisões e, por fim, respondendo ao mundo com inteligência" (INTEL, 2013, p. 3).

A origem desse grande volume de dados pode vir de variados lugares, como os milhões de websites, os milhões de cliques e compartilhamentos em redes sociais, dos bilhares de smartphones e dos diversos dispositivos com sensores, câmeras de monitoramento de cidades, bancos de dados, GPS dentre diversas outras fontes. Segundo Taurion (2013) cerca de 99% dos dados do mundo estão armazenados de forma digital.

> A crescente miniaturização da tecnologia bem como o aumento da sua capacidade de processamento e armazenamento permite a criação da Internet das Coisas, o que aumentará de forma exponencial a geração de dados. Taurion (2013, p. 30)

Na figura 1 é possível observar a enorme quantidade de dados geradas por minuto nas diversas redes e plataformas digitais, para dar conta de todo esse volume de dados é que surgiram as ferramentas específicas de *Big Data*.

Figura 1 – O que acontece na internet e um minuto



Fonte:LEWIS e CALLAHAN, 2018.

Para melhor contextualizar esse cenário foram definidos segundo Laney (2001) os três Vs do *Big Data*: volume (a grande quantidade de dados gerados), variedade (as variadas fontes de dados) e a velocidade (o processamento do grande volume de dados deve ser ágil na geração de informações de valor). Davenport (2014) expande essa definição para cinco Vs incluindo aos demais a veracidade (os dados gerados são verdadeiros?) e o valor (o dado obtido deve gerar valor para se obter informação útil).

Davenport (2014) refere-se ao Big Data como dados massivamente volumosos para caberem em simples servidores, extremamente desestruturados para se ajustarem a bancos de dados com base em linhas e colunas de tabelas relacionais, e continuamente fluídos para serem inseridos em estruturas estáticas de armazenagem.

2.2. *Big Data Analytics*

O termo *Big Data Analytics* segundo Cuzzocrea (2013) pode ser interpretado como procedimentos complexos que são executados em larga escala sobre grandes repositórios de dados, cujo objetivo é a extração de conhecimento útil mantido em tais repositórios. Em outras palavras, é a aplicação de técnicas analíticas avançadas a grandes conjuntos de dados.

Kimball e Ross (2013) afirmam que sua grandeza do *Big Data* não é a característica mais importante, mas sim, que este pode ser estruturado, semiestruturado, não estruturado, bruto e em muitos formatos diferentes, em alguns casos sendo totalmente diferentes do que os números escalares limpos e os textos que ficam armazenados em *data warehouses* durante anos. Segundo os autores muitos dados grandes não podem ser analisados com algo parecido com SQL.

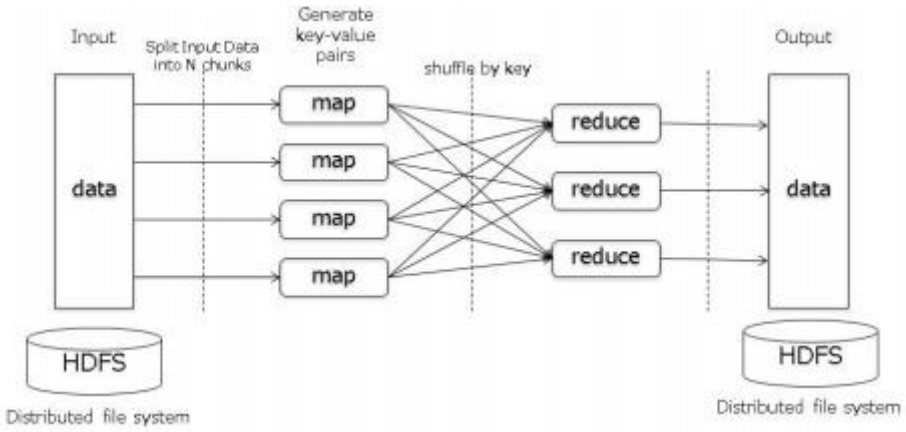
2.3. Hadoop

Hadoop é o mais conhecido framework de gestão de *Big Data*, mantido pela *Apache Software Foundation*, surgiu a partir do GoogleFile System (GFS) que foi publicado em 2003, este deu origem a outro trabalho da Google o MapReduce em 2004 (Cetax, 2017). Primeiramente foi denominado como Apache Nutch um subprojeto do Apache Lucene, em 2006 foi separado sob o nome Hadoop. É um software de código aberto e seu desenvolvimento é realizado por diversos contribuintes sendo o Yahoo! Inc. o maior deles além de maior usuário com cerca de 4500 nós em um cluster. No ano de 2008 deixou de ser um subprojeto e se transforma em um dos maiores projetos da Apache.

O Hadoop é uma estrutura que permite o processamento distribuído de grandes conjuntos de dados entre clusters de computadores usando modelos de programação simples

que foi projetado para expandir de servidores únicos para milhares de máquinas, cada uma oferecendo computação e armazenamento local. É baseado na arquitetura MapReduce que ao receber o arquivo os dados, reduz e os distribui nos *DataNodes* do HDFS que contém os blocos de dados replicados, conforme representado na figura 2. Ao invés de confiar no hardware, a própria biblioteca foi projetada para detectar e lidar com falhas na camada de aplicativos e fazer a redistribuição dos arquivos nos nós, cada um dos quais pode estar sujeito a falhas, dessa maneira é possível oferecer um serviço de alta disponibilidade no cluster.

Figura 2 – Representação de HDFS com MapReduce



Fonte: Honjo e Oikawa, 2013.

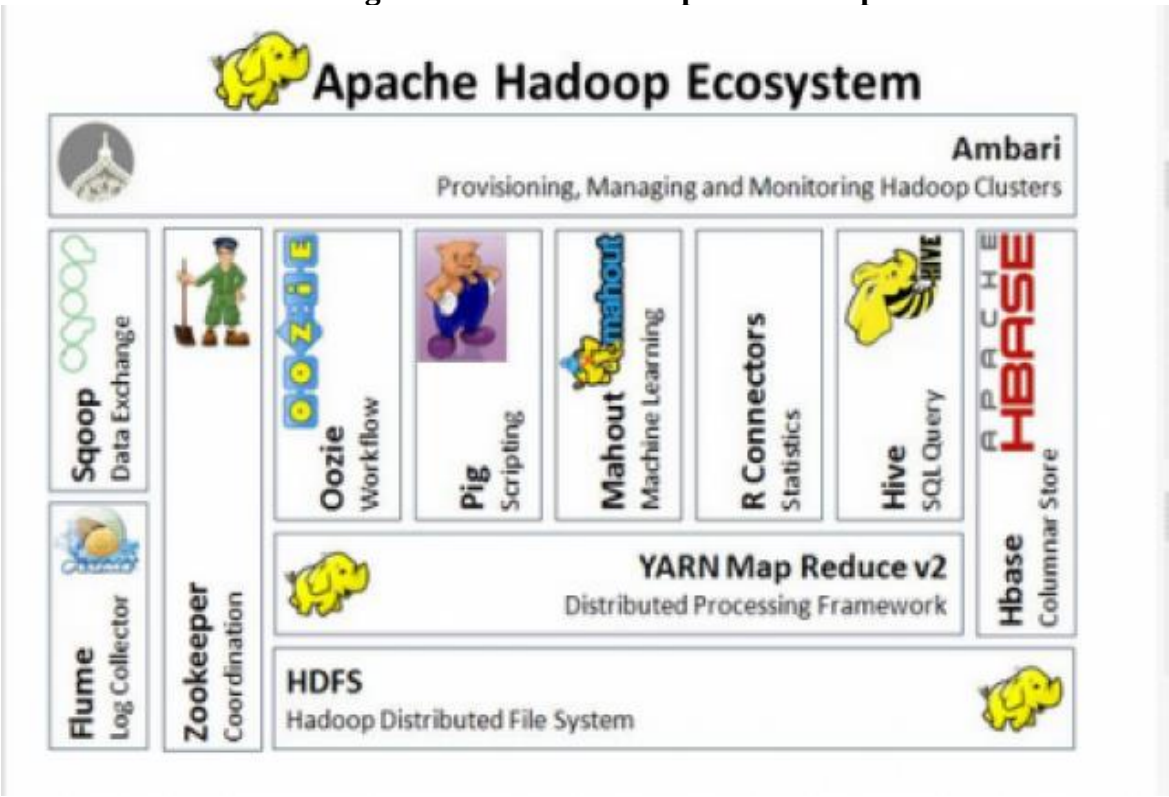
Compõem o framework Hadoop os seguintes módulos:

- Hadoop Common - Contém as bibliotecas e arquivos comuns necessários para todos os módulos Hadoop;
- HDFS - Hadoop Distributed File System ou sistema de arquivos distribuídos que surge para sanar a necessidade de se trabalhar com grandes quantidades de arquivos;
- MapReduce - é o sistema analítico do Hadoop que divide os arquivos e depois monta a separação dos dados em partições, mapeia as atividades em cada local, duplica em ambientes e depois faz as reduções;
- Yarn - é o gerenciador de recursos distribuídos do cluster.

2.4. Ecossistema Hadoop

Além do Hadoop existem outros projetos que estão diretamente relacionados ao Hadoop que é o principal deles aos quais formam um conjunto de ferramentas que permitem a coleta de dados de diversos lugares e em diversos formatos Hadoop. São diversas ferramentas usadas por empresas que utilizam o Hadoop que começaram a desenvolver outros componentes que se tornaram produtos do Apache Foundation e acabaram fazendo parte do seu ecossistema, conforme pode ser observado na Figura 3.

Figura 3 – Ecossistema Apache Hadoop



Fonte: http://blog.agro-know.com/?p=3810

Cada ferramenta tem uma função e trabalham juntamente com o Hadoop integrando o seu ecossistema.

3. PROCEDIMENTOS METODOLÓGICOS

O presente trabalho foi dividido em duas etapas para consecução da pesquisa, com a primeira etapa consistindo no desenvolvimento e montagem de um ecossistema de Big Data com computação distribuída em clusters utilizando o Hadoop e a segunda, em desenvolvimento,

focada na utilização do cluster como cliente para amostragem de *dashboards* utilizando ferramentas de *Business Intelligence*. As metodologias usadas para execução do trabalho foram levantamento bibliográfico e a implementação do cluster Hadoop.

a) Levantamento bibliográfico

O levantamento bibliográfico foi realizado através da abordagem de temas relacionados a analises de dados Big Data, ecossistemas Hadoop, entre outros assuntos da área de ciência de dados. A principal fonte de informação foram os sites Academia (https://www.academia.edu) e o Google Acadêmico (https:// scholar.google.com.br/) que são grandes repositórios de artigos e livros relacionados ao tema. Além disso foram realizadas buscas em sites de tutoriais sobre instalação e configuração do Hadoop, entre outros recursos.

Adotou-se o uso de dados abertos disponibilizados pela Câmara do Deputados Federal (https://www2.camara.leg.br/transparencia/) para execução do projeto.

b) Implementação do cluster Hadoop

Para a implementação do projeto foram utilizados o Hadoop e o Hive como *data warehose* além do MariaDB como banco de dados. Para tal foram configurados 1 computador pessoal e 3 notebooks com o sistema operacional CENTOS 7 e mais 3 máquinas virtuais com CENTOS7 para a montagem de um mini File System Hadoop (HDFS).

As etapas de execução projeto foram divididas em:

- Montagem e configuração das máquinas;
- Montagem do cluster Hadoop;
- Instalação do Apache Hive;
- Análise de dados usando ferramenta de *Business Intelligence* como cliente do cluster.

[1] Descrição de configuração disponíveis no Apêndice A

4. APRESENTAÇÃO E ANÁLISE DOS DADOS

O objetivo principal do projeto é a montagem de um cluster composto por máquinas físicas e virtuais para a construção de um sistema HDFS – Sistema de arquivos distribuídos do Hadoop, usando o Apache Hive como *data warehouse*, software que compõe o ecossistema Hadoop e por fim a conexão com a ferramenta de Business Intelligence PowerBI da Microsoft para apresentação dos dados, demonstrando na prática seus principais benefícios.

Para alcançar tal objetivo foram feitas as configurações de rede nas máquinas que

UNICEPLAC

ficaram com a seguinte configuração descritas na Tabela 1.

Tabela 1 – Configurações de rede

Nó Master	**Nó slave1**:	**Nó slave2**:	**Nó slave3**:
Rede interna:	IPv4 192.168.100.3	IPv4 192.168.100.4	IPv4 192.168.100.5
IPv4 192.168.100.2	Netmask 255.255.255.0	Netmask 255.255.255.0	Netmask 255.255.255.0
Netmask 255.255.255.0	Hostname "hadoop-slave1"	Hostname "hadoop-slave2"	Hostname "hadoop-slave3"
Hostname "hadoop-master"	**Nó slave4**:	**Nó slave5**:	
Rede externa:	IPv4 192.168.100.6	IPv4 192.168.100.7	
IPv4 DHCP Automático	Netmask 255.255.255.0	Netmask 255.255.255.0	
	Hostname "hadoop-slave4"	Hostname "hadoop-slave5"	

4.1. Distribuição de arquivos no Hadoop (HDFS)

Um arquivo no HDFS se divide em vários blocos e cada um é replicado dentro do cluster Hadoop. Um bloco no HDFS é um bloco de dados no sistema de arquivos subjacentes com um padrão tamanho de 64MB. O tamanho de um bloco pode ser estendido até 256 MB com base nas configurações feitas na instalação do Hadoop. O sistema de arquivos distribuídos (HDFS) armazena os dados e arquivos do cliente em metadados do sistema separadamente em servidores dedicados, nesse caso usaremos o Apache Hive para criação do *metastore*.

NameNode e *DataNode* são os dois componentes essenciais da arquitetura Hadoop HDFS.

NameNode: formado por Inodes que contém vários atributos, como permissões, registro de data e hora da modificação, cota de espaço em disco, cota de espaço para nome e horários de acesso. O *NameNode* mapeia a estrutura do sistema de arquivos inteira na memória e contém a lista de blocos que definem os metadados. Possui dois arquivos ‘fsimage’ e ‘edits’ que são usadas para persistência durante as reinicializações.

DataNode: gerencia o estado de um nó do HDFS e interage com os demais blocos. Um *DataNode* pode executar um uso intensivo da CPU, com trabalhos como análise semântica e de linguagem, estatísticas e tarefas de aprendizado de máquina (*machine learning*) e trabalhos intensivos de entrada e saída, importação de dados e exportação de dados, pesquisa, descompactação e indexação. Na inicialização, todo *DataNode* se conecta ao *NameNode* e verifica o ID do espaço para nome e a versão do software do *DataNode*. Se algum deles não corresponder, em última análise, o *DataNode* desliga automaticamente. Um *DataNode* verifica as réplicas de bloco em sua propriedade enviando um relatório de bloco para o *NameNode*.

Os dados do aplicativo são armazenados nos *DataNode*s e os metadados do sistema de arquivos é armazenado no *NameNode*. O HDFS replica o conteúdo do arquivo em vários

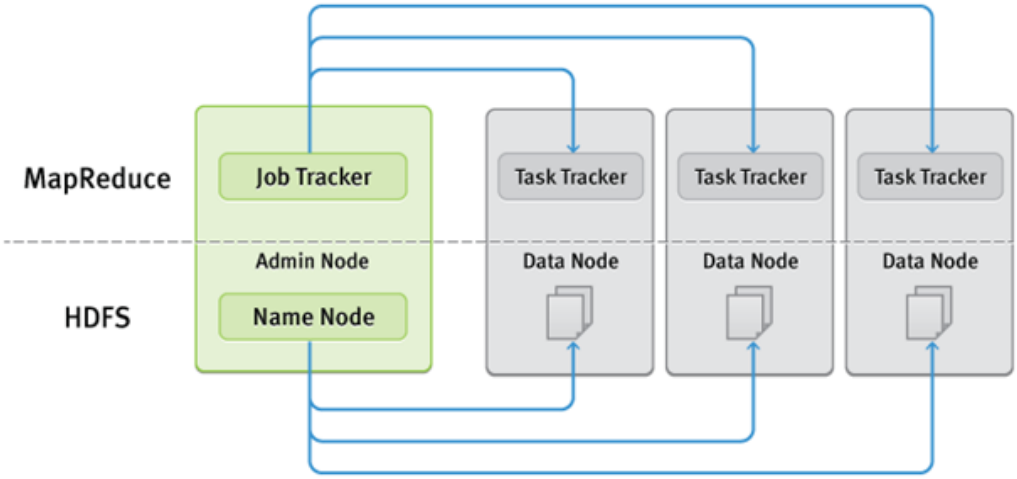
*DataNode*s com base no fator de replicação para garantir a confiabilidade dos dados. Para que o Hadoop possa ter desempenho eficiente, os discos rígidos devem ter alta taxa de transferência e boa velocidade de rede para gerenciar a transferência de dados e as replicações de bloco.

4.2. Hadoop atuando

O Hadoop foi desenvolvido para trabalhar com grandes volumes de dados, ou seja, para análises de *Big Data*, nesse sentido durante a evolução do projeto para execução e teste dos serviços foram usados dados abertos disponíveis no Portal brasileiro de dados abertos[7]. Os dados tratados foram salvos em *metastore* do Hive podendo assim serem consumidos.

O Hadoop segue um design de arquitetura *master slave*, para armazenamento de dados e processamento de dados distribuídos usando HDFS e MapReduce. O nó *master* serve para armazenamento de dados e para o processamento paralelo de dados usando o Hadoop MapReduce no HDFS, é no nó máster que se encontram o *NameNode* e o *JobTracker*. Os nós *slaves* na arquitetura Hadoop são as outras máquinas no cluster Hadoop que armazenam dados e realizam cálculos complexos. Todo nó escravo tem um *TaskTrack* e um *DataNode* que sincroniza os processos com o *NameNode* conforme pode ser observado na figura 4. A arquitetura Hadoop permite a implementação em nuvem ou máquina virtual, os nós master ou slave podem ser configurados em computação em nuvem como Azure da Microsoft e o AWS da Amazon.

Figura 4 – Arquitetura do Hadoop



Fonte: https://imasters.com.br/banco-de-dados/big-data-e-hadoop-o-que-e-tudo-isso

O funcionamento do Hadoop se dá pelas seguintes etapas

- Etapa 1: o usuário ou aplicativo envia uma tarefa ao Hadoop para processo

necessário, especificando os seguintes parâmetros:

a. A localização dos arquivos de entrada e saída no arquivo distribuído sistema.

b. O arquivo de configuração que contém a implementação de mapa e redução de funções.

c. A configuração do trabalho, definindo parâmetros diferentes específico para o trabalho.

Nesse caso a aplicação configurada para trabalhar com o Hadoop em nosso projeto é o Apache Hive.

- Etapa 2: O cliente da tarefa Hadoop envia a tarefa no *JobTracker* que assume a responsabilidade de distribuir a configuração para escravos, agendando tarefas e monitorando, o que fornece informações de status e diagnóstico para o cliente do trabalho.
- Etapa 3: Os Rastreadores de Tarefas em diferentes nós executam a tarefa conforme implementação do MapReduce e saída da redução A função é armazenada nos arquivos de saída no sistema de arquivos

4.3. Apresentação de dados

Os dados que serão inseridos no HDFS são de cota parlamentar da Camara Legislativa federal nos anos de 2016 e 2017, esta etapa está em desenvolvimento, a análise de dados se dará pelo uso do software Power BI que utilizará uma conexão ODBC. O PowerBI será como cliente consumindo os serviços (dados) do cluster Hadoop.

4.4. Benefícios do Hadoop

Segundo o BrandFinance[6](2019) as grandes empresas de tecnologia da informação lideram o ranking das 500 maiores marcas globais. O que as grandes marcas têm em comum é a grande quantidade de dados que geram e o grande desafio para essas empresas é gerir esse enorme volume de dados a fim de manter a sua performance no mercado. O Hadoop é uma solução que apresenta diversos benefícios as empresas que o adotam, os principais benefícios são:

[6]HAIGH, Richard. BrandFinance. Global 500 - 2019. Disponível em: https://brandirectory.com/rankings/global-500-2019

- Escalabilidade: A distribuição dos arquivos em e dados locais em nós de um cluster Hadoop permite que um armazenamento e processamento na escala de petabytes.
- Confiabilidade: Em vez de confiar no hardware para oferecer alta disponibilidade, o Hadoop foi projetado para detectar e lidar com falhas na camada de aplicativos, oferecendo um serviço altamente disponível em um cluster de computadores, cada um dos quais pode estar sujeito a falhas. Quando um nó falha no processamento é redirecionado para os nós restantes no cluster e os dados são automaticamente re-replicado em preparação para falhas de nó futuras.
- Flexibilidade: Os dados podem ser armazenados em qualquer formato, diferentemente dos tradicionais bancos de dados relacionais. É possível armazenar dados semi-estruturados ou não estruturados, e em seguida, analisar e aplicar esquema para os dados quando ler.
- Baixo Custo: Um dos grandes benefícios do Hadoop é que ele utiliza máquinas e redes convencionais, não sendo necessário investimentos em supercomputadores, sendo possível também ser utilizado os serviços disponíveis em nuvem, além disso o Hadoop é um software livre.

5. CONSIDERAÇÕES FINAIS

Conforme pode se verificar com o avanço tecnológico e a criação de novas tecnologias e a crescente geração continua de dados surgiu o conceito de *Big Data*, que segundo Santaché (2013) embora seja tratado por muitos como solução, em si é um problema, pela quantidade e diversidade de dados, o que reivindica o uso de ferramentas específicas de análises de *Big Data*. No entanto já existe diversas ferramentas que geram bons resultados nas análises de dados e pesquisas, gerando também bons resultados financeiros e operacionais as empresas.

O presente trabalho apresentou uma das ferramentas mais usadas e a mais famosa no mundo para processamento e manutenção de *Big Data*, o Hadoop que apresenta ser uma ferramenta eficiente confiável e de baixo custo. O Hadoop em si somente não agrega muito valor operacional, mas com o uso de outras ferramentas que compõem seu ecossistema, tais ferramentas que foram desenvolvidas para trabalharem juntas, em diversas camadas.

Existem diversos cases de sucesso de grandes empresas como a Amazon e o Facebook

que utilizam o Hadoop juntamente com as ferramentas de seu ecossistema para trabalhar com o grande volume de dados, vale ressaltar que o uso de ferramentas como o Hadoop é melhor aplicado em empresas que geram ou giram um grande volume de dados tal como as empresas citadas que trabalham com dados quase que em tempo real é recomendável para análises de *Big Data*.

REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR10520**: informação e documentação: citações em documentos: apresentação. Rio de Janeiro, 2002.

APACHE FOUNDATION. **HDFS ArchitectureGuide**. 2019. Disponível em: <https://hadoop.apache.org>. Acesso em: 14 out. 2019.

AVOYAN, Hovhannes. **Big Data e Hadoop:** o que é tudo isso?. 2014. Disponível em: <https://imasters.com.br/banco-de-dados/big-data-e-hadoop-o-que-e-tudo-isso>. Acesso em: 01 nov. 2019.

CÂMARA DOS DEPUTADOS: Dados Abertos - Cota Parlamentar. Brasilia: Câmara Leg, 2019. Disponível em: <https://www2.camara.leg.br/transparencia/cota-para-exercicio-da-atividade-parlamentar>. Acesso em: 15 nov. 2019.

CANAL TECH. ***Big Data*: os cinco Vs que todo mundo deveria saber.** Disponível em: <https://canaltech.com.br/big-data/Big-Data-os-cinco-Vs-que-todo-mundo-deveria-saber/>. Acesso em: 19 out. 2019.

CETAX. **HADOOP: O QUE É, CONCEITO E DEFINIÇÃO**. Disponível em: <https://www.cetax.com.br/blog/apache-hadoop/>. Acesso em: 4 out. 2019.

CUZZOCREA, Alfredo. Analytics over Big Data: Exploring the Convergence of DataWarehousing, OLAP and Data-Intensive Cloud Infrastructures. **2013 Ieee 37th Annual Computer Software And Applications Conference**, [s.l.], p.481-483, jul. 2013. IEEE. http://dx.doi.org/10.1109/compsac.2013.152.

DAVENPORT, Thomas H. **Big Data at Work: Dispelling the Myths, Uncovering the Opportunities.** Cambridge: Harvard Business Review Press, 2014.

DITTRICH, Jens; QUIANÉ-RUIZ, Jorge-arnulfo. Efficient big data processing in Hadoop MapReduce. **Proceedings Of The Vldb Endowment**, [s.l.], v. 5, n. 12, p.2014-2015, 1 ago. 2012. VLDB Endowment. DOI: http://dx.doi.org/10.14778/2367502.2367562.

ENOMURA, Bianca Y. ***Big Data***: A era dos grandes dados já chegou. 2014. 22 f. TCC (Graduação) - Curso de Jornalismo, Departamento de Jornalismo, Universidade Federal de Santa Catarina, Florianópolis, 2014. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/133419/BIG%20DATA%20-%20Superinteressante.pdf?sequence=1>. Acesso em: 4 out. 2019.

GALDINO, Natanael. ***Big Data*: Ferramentas e Aplicabilidade**. Disponível em: <https://www.aedb.br/seget/arquivos/artigos16/472427.pdf>. Acesso em: 06 out. 2019

GARTNER. ***Big Data***. 2012. Disponível em: <https://www.gartner.com/it-glossary/big-data/>. Acesso em: 13 out. 2019.

GASPARINI, Vinicius. **Estudo da Madrugada, desvendando o ecossistema do Apache Hadoop**. Disponível em: <https://medium.com/vinicius-gasparini/estudo-da-madruga-desvendando-o-ecossistema-do-apache-hadoop-9c64a556fbf7>. Acesso em: 19 out. 2019

HAIGH, Richard. BrandFinance. Global 500 - 2019. Disponível em: <https://brandirectory.com/rankings/global-500-2019>. Acesso em 12 out 2019

HONJO, Toshimori; OIKAWA, Kazuki. Hardware acceleration of Hadoop MapReduce. **2013 Ieee International Conference On Big Data**, [s.l.], p.118-124, out. 2013. IEEE. DOI: http://dx.doi.org/10.1109/bigdata.2013.6691562.

KIMBAL, Ralph; ROSS, Margy. **The Data Warehouse Toolkit:** The Definitive Guide to Dimensional Modeling. 3. ed. Indianopolis: John Wiley & Sons, 2013.

MACHADO, Felipe N. R. ***Big Data* O Futuro dos Dados e Aplicações**. São Paulo: Saraiva, 2018. 224 p.

TAURION, Cezar. ***Big Data***. Rio de Janeiro: Brasport, 2013.

CALDAS, M. S.; CLAUDINO SILVA, E. C. Fundamentos e aplicação do Big Data: como tratar informações em uma sociedade de yottabytes. **Bibliotecas Universitárias: pesquisas, experiências e perspectivas**, v. 3, n. 1, 13 abr. 2016.

S., Ashlesha; M., R.. A Review of Hadoop Ecosystem for BigData. **International Journal Of Computer Applications**, [s.l.], v. 180, n. 14, p.35-40, 17 jan. 2018. Foundation of Computer Science. http://dx.doi.org/10.5120/ijca2018916273.

SANTANCHÈ, A. - **NoSQL e *Big Data*** - Aula 27 - Bancos de Dados 2015.2
Disponível em: <https://www.youtube.com/watch?v=-a2pyU0uhww>. Acesso em: 10 out. 2019.

SHIM, Kyuseok. MapReduce algorithms for big data analysis. **Proceedings Of The Vldb Endowment**, [s.l.], v. 5, n. 12, p.2016-2017, 1 ago. 2012. VLDB Endowment. DOI: http://dx.doi.org/10.14778/2367502.2367563.

VICTORINO, Marcio de Carvalho et al. UMA PROPOSTA DE ECOSSISTEMA DE BIG DATA PARA A ANÁLISE DE DADOS ABERTOS GOVERNAMENTAIS CONE CTADOS. **Inf. & Soc.:est,** João Pessoa, v. 27, n. 1, p.225-242, jan. 2017. Quadrimestral. DOI: http://dx.doi.org/10.22478/ufpb.1809-4783.2017v27n1.29299

APÊNDICE A – CONFIGURAÇÃO DO CLUSTER HADOOP E HIVE

UNICEPLAC

[1]

Montagem e configuração das máquinas:

São necessárias pelo menos três máquinas físicas ou virtuais para que o cluster Hadoop funcione, podendo esse número aumentar, quanto mais nós em um cluster mais rápido ele será. Os pré-requisitos mínimos necessários para construção do ambiente Hadoop são:

- 2 GB de memória RAM,
- 20 GB de espaço livre no HD,
- 2 placas de rede,
- Sistema Operacional CENTOS 7, com acesso como root nas máquinas, usuário "hadoop" nas máquinas,
- Conexão com a internet para fazer os downloads dos pacotes necessários,
- VirtualBox no caso de máquinas virtuais com o CENTOS 7.

Na etapa de instalação, o primeiro passo foi a configuração das placas de rede que possibilitaram a conexão com a internet e com as demais máquinas do cluster (*DataNode*s). Uma placa configurada em modo Bridge para estabelecer conexão com a internet e uma segunda em modo de "rede interna", placa que será responsável em fazer a conexão entre as máquinas.

Foi necessário também desativação do SELINUX (mecanismo de segurança Mac), responsável por regras de segurança em arquivos e processos em sistemas Linux. Desativação do firewall para que não sejam encontrados problemas relacionados a portas bloqueadas durante o trabalho das máquinas em conjunto e a instalação do Java JDK (Java Development Kit).

Comunicação:

Para comunicação entre as máquinas foi necessário editar o arquivo "hosts" em cada uma das máquinas e inserir os endereços de IP das demais, com isso elas conseguiram acesso entre sí. Configurar as variáveis de ambiente JDK (Java Development Kit) e apontar o caminho da variável Hadoop, isso é possível com a edição do arquivo "~/.bashrc".

Foi feita a configuração do SSH com a chave de autenticação de chave RSA (Autenticação Rivest Shamir Adleman), o que tornou possível o gerenciamento de atividades de forma local ou remota. Para a conexão SSH foi necessário a edição do arquivo "authorized_keys". Foi necessário duplicar a linha da chave pública que já está presente dentro do arquivo "authorized_keys" obtendo linhas idênticas. Feito isso foram feitas mudanças no nome dos hosts que está no final de cada linha para o nome de cada nó do cluster.

Após realizados os procedimentos de conexão SSH, continuamos copiando as chaves públicas dentro de cada nó inclusive no "localhost" de cada máquina que irão formar o cluster. Ao fim foram alteradas as permissões em todos os hosts do cluster. Esses comandos foram realizados no host "hadoop-master".

Instalação e configuração do Hadoop

Após a instalação e configuração do sistema operacional nas máquinas, pode-se avançar para a instalação do Apache Hadoop.

Primeiro passo foi realizar o download do Apache Hadoop, após o download, configuração dos arquivos "core-site.xml", "hdfs-site.xml", "mapred-site.xml.template" e criação de uma cópia do último arquivo com o nome "mapred-site.xml". Foi feita a edição do arquivo de configuração "yarn-site.xml","hadoop-env.sh" e a criação dos arquivos “master” em que foi inserido o nome “hadoop-master” e “slave” em que foram inseridos os nomes dos nós slaves.

Foi necessário a criação de diretórios para conter arquivos de log e arquivos temporários e diretórios para conter os "*NameNode*" e "*DataNode*".

Após a configuração e a criação dos diretórios, foi distribuída uma cópia já configurada da máquina “hadoop-master” para todos os nós disponíveis. Feito isso preparamos o cluster HDFS para que começar a coletar os dados e por fim iniciado o serviço do yarn (para monitoramento de recursos).

Instalação e configuração do Hive

Os pré-requisitos necessários para construção do ambiente Hadoop são:

- Uma cluster Hadoop
- Apache Hive.
- Banco de dados MariaDB (Mysql)

Na etapa de instalação, o primeiro passo foi o download e a instalação do Apache Hive e a instalação do server MariaDB, habilita-lo e configura-lo paraser iniciado durante o boot e configurar o Mysql Server.

Foi configurado no banco de dados MariaDBo usuário e as tabelas do Hive no MySql, a criação do banco foi feita para ser usado apenas como armazenamento do metastore.

Para configuração do Hive foi necessário a criação do arquivo "hive-site.xml" e inserção das tags de configuração. Foi feita a cópiado arquivo hive-env.sh.template" para o "hive-env.sh", e alteração "hive-env.sh" indicando o caminho de configuração do Hive.

Após realizadas as configurações foi feita a cópia da biblioteca para fazer a conexão entre o Hive e o Mysql. Feito isso serviços do Hive Metastore e HiveServe podem ser inicializados.

Agradecimentos

Agradecemos a Deus por ter no ajudado chegar até aqui.

E a todos que de alguma forma contribuíram para o crescimento de nosso trabalho.