ARTIGO CIENTÍFICO
ÁREA DE CONCENTRAÇÃO

CIÊNCIA E TECNOLOGIA

HADOOP: UMA NECESSIDADE PARA ANALISTAS

Eliezer de Souza Batista Junior[1], Rôber Yamashita[2]
Mestre em Ciências Militares[1], Doutorando em Business Administration[2]

RESUMO: *Big Data* é um termo que está em voga atualmente. A necessidade de realizar pesquisas em uma grande quantidade de dados, de diferentes formatos, resultando em produtos rápidos, assertivos e com grau de valor que os adversários não possuem é alta em um ambiente de competição. Para tanto, analistas necessitam de ferramentas que provém esses requisitos para que seus chefes possam, de melhor forma, realizar o processo da tomada de decisão. O Hadoop é uma dessas ferramentas e o seu uso será discutido como principal objetivo desse artigo. Outros objetivos específicos são caracterizar *Big Data* e Hadoop e mostrar os desafios de implementação da funcionalidade.

Palavras-chave: Big Data, Dados, Decisão, Ferramentas, Hadoop

INTRODUÇÃO

Com a Internet das coisas (IoT), houve um grande aumento na quantidade de dispositivos com capacidade de gerar dados [Cozza et al 2011]. Os dados provenientes desses dispositivos são armazenados em diferentes repositórios, possuindo suas peculiaridades de formatação, segurança e rede. Nesse ínterim mídias sociais, blogs, anúncios, vendas e transações bancárias são alguns exemplos de onde os dados são gerados, podendo até chegar em grandes planejamentos de cidades conectadas, ou *smart cities* [Moreno-Cano, 2015].

A integração desses dados é algo que, naturalmente, se faz necessário. Os sistemas não trabalham mais como "ilhas", como ocorria no passado. Atualmente, o resultado do processamento de um software pode ser o início de um outro processo. Por exemplo, a análise de reclamações bancárias no portal "reclame-aqui" é o início do processo de tratamento do problema pela parte jurídica de uma determinada empresa.

Coletar esses dados é uma tarefa extremamente difícil, pois não está se falando apenas dos dados contidos em repositórios de aplicações corporativas internas, mas também das interações que essas realizam no espaço virtual e com outros repositórios externos [Tan et al. 2013]. Ferramentas capazes de extrair, transformar e carregar dados em um repositório são primordiais para empresas que querem possuir vantagens competitivas.

Infelizmente, a velocidade de crescimento de dados geralmente não é proporcional com o crescimento de hardwares, com exceção de empresas que podem bancar por tal situação. Os preços de equipamentos de ponta-de-linha e suas respectivas manutenções estão cada vez mais caros para se adquirir [Tulloch, 2019]. A vertente econômica que diz: "necessidades ilimitadas para recursos escassos" é uma premissa em ambientes de centro de processamento de dados ou em *datacenters*. Equipamentos storages geralmente têm a necessidade de realizar upgrades com menos de 05 (cinco) anos por conta do volume de dados gerados, ocasionando superlotação.

Outro fato que causa bastante impacto é que profissionais de Tecnologia da Informação (TI) necessitam de tecnologias adequadas. Do ponto de vista de armazenamento, os atuais Sistemas Gerenciais de Banco de Dados (SGDB) disponíveis comercialmente não são capazes de lidar com grandes volumes de dados. A análise de grandes volumes requer SGDBs especializados capazes de processar dados estruturados e não estruturados, distribuindo dados a fim de escalar grandes tamanhos [Begoli e Horey, 2012].

A velocidade do dado é outro ponto que causa bastante impacto nos sistemas de

TI. Dependendo do local a ser coletado, muitos profissionais preferem trabalhar apenas com extratos (amostras) dos dados para que sejam analisados, o que não permite que todos os detalhes daquele conjunto sejam observados [DiFranzo et al. 2013].

Os dados que trafegam na rede mundial de computadores possuem diferentes formatos como imagens, vídeos, textos, agrupamento de dados, dentre outros. Esses formam um grande conglomerado que não possui estruturação fixa. Segundo Chen (2001), pode-se perceber que grande parte da web é composta por dados não estruturados.

Com os dados corretos e apresentados de forma correta, as fases de observação e orientação do ciclo de tomada de decisão OODA fica mais fidedigno para a próxima ação que é a decisão. Portanto, a tomada de decisão é diretamente impactada por conhecimentos extraídos de diversas bases de dados. [Hazen e Jones-Farmer, 2014]

Com todos os problemas relatados nessa seção, verifica-se a necessidade de ferramentas para dar uma solução à demanda de tratamento do *Big Data*. Os próximos capítulos abordarão sobre o *Big Data* em uma metodologia que se inicia em conceitos mais amplos até chegar nos processos mais específicos de casos de implementação. Com isso, abordar-se-á: *Big Data*, casos de sucesso, Hadoop, possíveis locais de implementação e desafios.

2 *BIG DATA*

Por muito tempo, as empresas não trataram os dados que são trafegados fora e dentro de seus sistemas. Em outras palavras, não davam valor aos dados que possuíam. Quando a primeira empresa conseguiu o devido tratamento, gerou uma grande vantagem competitiva em relação às demais empresas concorrentes e conquistou a ponta dos negócios. O modelo foi replicado pelas concorrentes e a ferramenta que significava a liderança do mercado passou a ser a correta análise.

Segundo Dean e Ghemawat (2004), a primeira empresa que processou grandes quantidades de dados em paralelo, dividindo o trabalho em conjuntos de tarefas independentes, foi a Google, em 2004, com o desenvolvimento do *MapReduce*. Em 2005, o sistema foi melhorado pela Yahoo, sendo que alguns autores creditam a criação do *Big Data* a essa empresa, como é o caso da Chede (2011).

A definição de *Big Data* possui várias versões. Segundo o Gartner (2012), *Big Data*, em geral, é definido como “ativos de informações de alto volume, alta velocidade e/ou alta variedade que exigem formas inovadoras e econômicas de processamento de informações que permitem uma visão aprimorada, tomada de decisão e automação de processos”. A Intel (2014), define como sendo:

> Um conjunto de dados extremamente amplos e que, por este motivo, necessitam de ferramentas especialmente preparadas para lidar com grandes volumes, de forma que toda e qualquer informação processada por esses meios possa ser encontrada, analisada e aproveitada em tempo hábil: O valor 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.

Na literatura pesquisada sobre o assunto, as definições disponíveis tratam o tema se utilizando de características. Doug Laney (2001) utiliza a notação 3V para designar as principais características que são:

- Volume: organizações coletam dados de grande variedade de fontes, incluindo transações comerciais, redes sociais e informações de sensores ou dados transmitidos entre máquinas. Novas tecnologias têm possibilitado realizar esse requisito de armazenamento que era muito difícil de ser cumprido antes do aparecimento do *Big Data*;

- Velocidade: a velocidade dos dados aumentou substancialmente nos últimos anos. Tratá-los em tempo hábil sem perder qualquer tipo de informação também é um desafio;

c. variedade: dados estruturados são complicados de serem coletados e armazenados em um repositório comum. Mesmo possuindo padrões, ainda assim necessita-se realizar normalização de tabelas para que se tenha um entendimento do que se está tratando. O problema é potencializado quando são adicionados textos, e-mails, vídeos, áudios e sons sem padrão. Como não há padronizações, denomina-se esses de dados não estruturados. Ainda há os dados que são semi-estruturados, caracterizados por possuírem padrões heterogêneos, e portanto, são difíceis de serem identificados por seguirem diversos padrões. Com isso, aumenta-se a faixa de atuação do *Big Data*;

Há uma linha de raciocínio mais abrangente, como relata Gomes (2018), que acrescentam outras duas características em complemento ao 3V, formando 5V:

- Veracidade: esta dimensão inclui a consistência dos dados (confiabilidade estática) e a confiabilidade dos dados definida pelo número de fatores (incluindo a origem dos dados, métodos de coleta, processamento e infraestrutura confiável).

- Valor: definida pelo valor agregado que o dado coletado pode trazer para um processo, atividade ou hipótese.

FIGURA 1 5V do *Big Data*

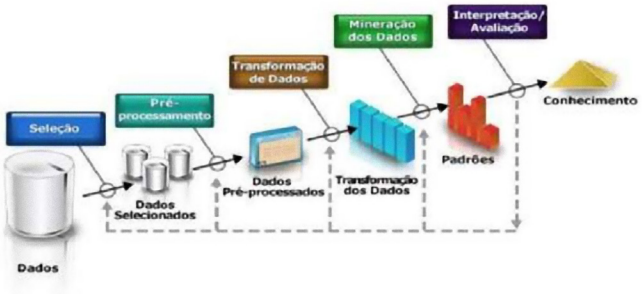
Fonte: Laney (2001)

O principal objetivo do *Big Data* é produzir conhecimento por meio de bancos de dados. Para atingir esse alvo, Fayyad (1996) propôs uma metodologia de "Descoberta de conhecimento em Banco de Dados (KDD):

> KDD é um processo, de várias etapas não trivial, interativo e interativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados.

A metodologia é dividida em 05 (cinco) etapas, conforme mostra a figura 2.

FIGURA 2 Etapas do *Big Data*



Fonte: Fayyad (1996)

Ainda segundo Fayyad (1996), as cinco etapas são as seguintes:

- Seleção: é o momento em que é realizada a coleta dos dados que serão avaliados no KDD. A qualidade do dado levantado nessa fase pode impactar diretamente o resultado;

- Pré-processamento: após coleta dos dados realizada na fase anterior, é realizada uma filtragem para retirada de caracteres indesejados, ruídos ou informações incompletas. Esses erros ocorrem, geralmente, em bases de dados heterogêneas que não possuem devido tratamento de dados;

- Transformação: objetiva adequar os dados em uma estrutura e formatação necessária conforme exigido no algoritmo de mineração de dados (próxima fase). Na transformação, ocorre a conversão de tipos de dados;

- Mineração: processo automático ou semi-automático que utiliza algoritmos e visa explorar e analisar grandes bases de dados, objetivando encontrar novos padrões e regras úteis e compreensíveis para o analista; e

- Exibição dos resultados: são realizadas seleções e ordenações das descobertas de interesse, gerando relatórios de resultados.

Segundo o Westcon (2019), caso seja implementado de forma correta, o *Big Data* pode trazer vários benefícios como por exemplo melhoria na tomada de decisão, identificação de padrões, acompanhamento da concorrência, melhores estratégias de marketing, relacionamento com o cliente, otimização de processos internos, gerenciamento de risco e melhoria da cibersegurança.

3 HADOOP

Uma das formas de implementação do *Big Data* é por meio da utilização do Apache Hadoop. Conforme o próprio site, o Apache Hadoop é um projeto que desenvolve software de código aberto para computação distribuída, confiável e escalável. Sua biblioteca é uma estrutura que permite o processamento distribuído de grandes conjuntos de dados entre *clusters* (conjunto de computadores que compartilham um sistema de arquivos) de computadores, usando modelos de programação simples. Ele foi projetado para expandir de servidores único para milhares de máquinas, cada uma oferecendo computação e armazenamento local. Em vez de confiar no hardware para oferecer alta disponibilidade, a própria biblioteca foi projetada para detectar e lidar com falhas na camada de aplicativos, oferecendo um serviço altamente disponível em cluster de computadores, cada um dos quais pode estar sujeito a falhas. (HADOOP, 2019)

Os benefícios da utilização do Hadoop iniciam por ser um software de código aberto, ou seja, não há necessidade de pagamento. Segundo o site SAS (2019), apesar disso, versões comerciais (chamadas de “distros”) têm sido lançadas. Essas distros são pagas e o usuário recebe capacidades adicionais relacionadas à segurança, governança, SQL, consoles de gestão/administração, treinamento, documentação e outros serviços. Os distros mais populares são Cloudera, Hortonworks, MapR, IBM BigInsights e PivotalHD.

Segundo CETAX, outros benefícios do uso do Hadoop são:

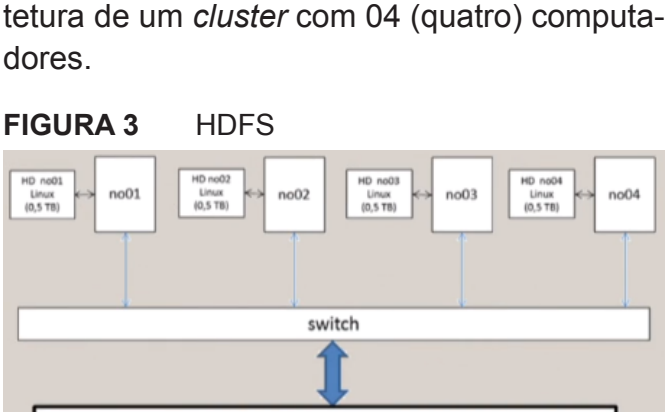
- Escalabilidade e desempenho: permite armazenar, gerenciar, processar e analisar dados em escala de petabytes;

- Confiabilidade: a ferramenta é resistente à falhas por conta de replicação; e

- Flexibilidade: há possibilidade de armazenar os dados em qualquer formato.

O Hadoop é um sistema que é instalado em plataforma operacional Linux. Verificou-se que a instalação no CentOS possui menos erros de instalação que outras distribuições. O sistema de arquivos utilizado pelo Hadoop é o HDFS (*Hadoop Distributed FileSystem*, ou Sistema de Arquivos Distribuído do Hadoop). Foi projetado para armazenar arquivos muito grandes, escrever uma vez e ler muitas vezes e funcionar com *commodity hardware* que são definidas como máquinas simples com preço acessível, mas que não significa baixa capacidade. Cada computador que constitui o *cluster* terá uma pequena partição para armazenar o sistema operacional. A figura 3, mostra a arquitetura de um *cluster* com 04 (quatro) computadores.

FIGURA 3 HDFS



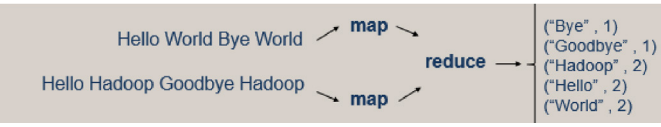
Fonte: Estácio de Sá (2019)

Enquanto sistemas operacionais normais operam com blocos de disco de 1KB (1024 bytes), o HDFS trabalha, por padrão,

com 64 MB. Entretanto, esse valor pode chegar à 128 MB. (Estácio de Sá, 2019)

O principal processo executado é o MapReduce que trabalha com mapeamento de tarefas que gera um par de chave e valor e depois executa com a redução do tamanho de um arquivo, tendo como base o par de chave e valor. Um exemplo de como esse processo é realizado é mostrado na figura 4.

FIGURA 4 MapReduce



Fonte: Estácio de Sá (2019)

A versão mais recente do Hadoop possui um módulo chamado de YARN (*Yet Another Resource Negotiator* – ainda outro negociador de recursos) que é constituído por uma estrutura de agendamento de tarefas e gestão de recursos de *clusters* (Pinto, 2015). Sua arquitetura é constituída dos seguintes componentes:

- *ResourceManager*: agente que arbitra os recursos (CPU, memória, disco, rede etc.) entre todas aplicações do sistema. É constituído de *scheduler* e *ApplicationManager*;

- *Scheduler:* aloca recursos às várias aplicações em execução;

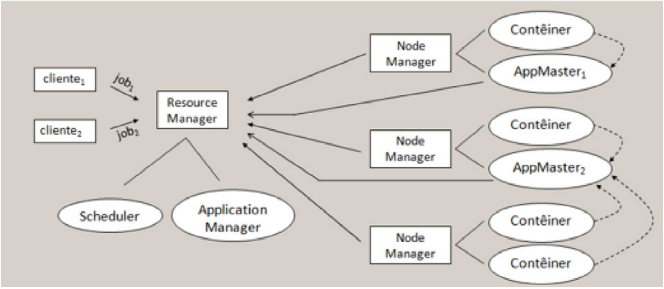
- *ApplicationManager*: negocia com os contêineres para execução da aplicação de tarefas. Também é responsável por monitorar erros;

- *NodeManager*: Responsável pelos contêineres, monitorando os recursos utilizados por eles, e fornecendo relatórios desses usos para o *Scheduler*; e

- *ApplicationMaster*: encarregado de negociar contêineres do *Scheduler*. Trabalha com o *NodeManager* para executar tarefas.

A figura a seguir mostra como uma tarefa solicitada por um cliente é desenvolvida no Hadoop versão 2.

FIGURA 5 Arquitetura do Yarn



Fonte: Estácio de Sá (2019)

Segundo SAS, outros componentes de software podem ser executados sobre ou junto com o Hadoop. Esses softwares formam o ecossistema Hadoop. A utilização não é obrigatória e depende dos requisitos do cliente. Alguns exemplos são:

- Ambari: interface web para gerir, configurar e testar serviços e componentes do Hadoop;

- Avro: software para troca de dados serializados e serviços;

- Cassandra: banco de dados distribuído;

- Drill: interface para smartphones e *internet banking*;

- Flume: coleta, agrega e coloca grandes quantidades de fluxos de dados em HDFS;

- HBase: banco de dados distribuído e não relacional. As tabeles podem servir como entradas ou saídas para trabalhos e *MapReduce*;

- HCatalog: gestor de tabelas e armazenamento que ajuda usuários a acessar e compartilhar dados;

- Hive: *Data Warehouse* e uma linguagem de consulta semelhante ao SQL que apresenta dados na forma de tabelas;

- Mahout: trabalha com algoritmo de aprendizado de máquina e biblioteca de mineração de dados;

- Oozie: organizador de tarefas;

- Pig: plataforma para manipular dados armazenados em HDFS que inclui um compilador para programas *MapReduce* e uma linguagem de alto nível, chamada de *Pig Latin*.

- Solr: ferramenta escalável que inclui indexação, confiabilidade, configuração central, tolerância a falhas e recuperação;

- Spark: estrutura de computação em cluster com inteligência analítica *in-memory*;

- Sqoop: mecanismo de conexão e transferência que movimenta dados entre Hadoop e bases de dados relacionais; e

- Zookeeper: coordena o processamento distribuído.

A figura 6 mostra o atual ecossistema Hadoop com alguns softwares que foram necessários, nesse caso.

FIGURA 6 Exemplo de arquitetura do sistema Hadoop

Fonte: Data-Flair (2019)

4 LOCAIS DE IMPLEMENTAÇÃO E CASOS DE SUCESSO

Atualmente, um dos fatores que permite a vantagem competitiva é a análise de informações de forma fidedigna. Várias empresas necessitam do entendimento do mundo que a cerca para tomada de decisão.

Um dos locais que mais utilizam o *Big Data* é o comércio varejista. Necessidade de conhecer como, quando, por que, o que um cliente compra são exemplos de perguntas que melhoram o rendimento das vendas. Chin et al (2016) descreve o caso de estudo da Amazon. A Amazon era uma empresa que vendia algumas quantidades de livros até os anos 2000. Em 2004, após a percepção do comportamento de seus clientes, a Amazon expandiu os seus negócios e seus serviços. Chegou ao ponto de disponibilizar a plataforma de *Big Data* para que outras empresas pudessem realizar a mesma operação.

Outro local que é de interesse são hospitais. Alguns conhecimentos podem trazer respostas para demandas que não foram completamente solucionadas. Por exemplo, ao comparar o local onde um paciente vive e sua doença com outras pessoas, poder-se-ia chegar a uma conclusão que a localidade é endêmica. Oliveira (2014) mostra que grande parte dos hospitais norte-americanos usam o *Big Data*, como é o caso do *Reliant Medical Group*, em Boston. O hospital utiliza acessórios da Apple e Google, como *smartwatches*, para servirem como sensores para monitorar saúde de cardíacos, por exemplo.

As operações bancárias também necessitam de conhecimentos que podem ser providos por sistemas de *Big Data*. Saber se uma operação foi realmente executada por um cliente, baseado em seu perfil, é um exemplo que Prakash (2015) descreve. O Hadoop é uma das ferramentas mais aplicadas nesse ramo, principalmente para uso de linguagens que não são baseadas apenas em SQL. Além disso, milhares de transações bancárias são realizadas por dia. Fazer um sistema resiliente a erros é um grande desafio.

Dentro da administração de cidades também é possível perceber a utilização de *Big Data*. Segundo Cheng (2015), o software chamado de CiDAP (*City Data and Analytics Platform*) foi capaz de elevar a cidade de Santander à condição de cidade inteligente que consegue responder a questionamentos como relacionados à energia, coleta de lixo, serviços prestados pela prefeitura e trafegabilidade. Há de se ressaltar que o referido projeto possui uma parte onde foi implementado o Hadoop para absorver dados não estruturados.

O Sistema Jurídico no Brasil já aplica sistemas de *Big Data*. Segundo Melo (2017), o *Big Data* no Judiciário mostra como dados impactam nos negócios dos tribunais (distribuição de justiça) e como a presidência da corte pode tomar ações antecipadas, especialmente para conferir mais celeridade aos Tribunais de Justiça.

No mundo militar, há alta demanda para com dados e informações que exigem a mais alta segurança. Andrejevic e Gates (2014) abordam que as principiais agências de inteligência norte-americanas (NSA, CIA e forças militares) trabalham com *Big Data*. Em outras palavras, não é apenas um requisito para melhorar as capacidades de defesa, mas também de segurança. Silva (2016) destaca o uso de *Big Data* ao dizer que "suas possíveis aplicações militares vêm sendo exploradas especialmente no que se refere à modelagem e à simulação, assim como em atividades de teste e avaliação". Várias são as possibilidades de se utilizar o *Big Data* e Hadoop em operações militares.

Vários outros casos de estudo de *Big Data* existem e são possíveis de utilização. Os exemplos supracitados dão uma noção ao leitor de como a tecnologia é poderosa e pode trazer benefícios.

5 DESAFIOS

A implementação do Hadoop não é uma atividade simples. A primeira definição que se deve realizar é saber o porquê da implementação do Hadoop. Muitos demandantes não sabem o que podem extrair com essa poderosa ferramenta. Após um processo de análise que pode ser muito custoso, o decisor pode entender que aquela informação extraída não é de interesse.

Após decidir sobre a implementação do Hadoop e levantado os requisitos de utilização, faz-se necessária a definição de arquitetura do sistema. Conforme dito no capítulo 3 (Hadoop), não há necessidade de implementar todas as ferramentas do ecossistema, mas há necessidade de "pinçar" as ferramentas que atenderão às necessidades elencadas para que, em um momento posterior, haja a sincronização.

Em termos de necessidades, muitos comerciantes dirão que *Business Inteligence* (BI) possui as mesmas capacidades que o *Big Data*. Entretanto, segundo Fiveacts (2014), tal assertiva é falsa. "Enquanto as soluções de *Big Data* servem para minerar dados de forma mais precisa, a ferramentas de BI analisam e condensam informações para tomada de decisões. É uma dupla infalível para a competitividade da empresa". Verifica-se que BI é algo mais personalizado, enquanto *Big Data* é mais amplo. Os dois não se substituem, se complementam. Portanto, de acordo com os requisitos postos em cena, o BI poderá ser uma opção a ser considerada.

Segundo Ferreira e Picchetti (2018), em empresas com muito tempo de operação, um dos grandes desafios é a mudança de paradigmas relacionado aos recursos humanos. Para que o sistema funcione com efetividade, há necessidade do uso de dados fidedignos. Por exemplo, se os funcionários de um determinado hospital não lançam dados ou dados incorretos referentes a procedimentos médicos, o sistema produzirá informações erradas e, por consequência, entrará em descrédito. Há necessidade de realizar trabalhos de conscientização dos funcionários para que o Hadoop funcione corretamente.

A implementação em si é outro desafio que ocorre. O Hadoop é baseado em Java 1.7 que já está desatualizado. Isso ocorre porque os javas novos ainda não foram personalizados às demandas que o Hadoop possui. Dessa forma, na própria instalação da ferramenta, há necessidade de instalação do Java 1.7 de forma prévia. Isso traz problemas de segurança.

Outro desafio é achar os softwares que constituem o ecossistema do Hadoop de forma compatível. Gomes (2017) relata que em sua experiência teve problemas com o Kettle (versão 6.0) e o Apache Cassandra (versão 3.4).

Assim como o Java, houve a necessidade de buscar uma versão mais antiga do Cassandra (2.0.17) para que a sincronização fosse realizada.

Outro problema é em relação à hardware. Segundo Hortonworks (2019), os softwares do ecossistema Hadoop necessitam de muita memória disponível e dedicada. A versão do Hortomworks que é utilizada para testes, treinamentos e aprendizado solicita 8 GB de memória RAM, minimamente, para que as funcionalidades sejam utilizadas com efetividade. Em outras palavras, em lap tops normais (possuem cerca de 8GB) não é possível o uso. Em usos profissionais, esses requisitos aumentam de acordo com o *commodity hardware* que está sendo utilizado.

FIGURA 6 Quantidade de vulnerabilidades, expressadas por CVE

Fonte: Bhathal e Singh (2019)

Até o momento no ano de 2019, foram descobertas 49 vulnerabilidades. Dessa forma, colocar um ativo de Tecnologia da Informação conectado diretamente à Internet não é um bom procedimento. Há necessidade de inserir instrumentos de segurança, como IPS, IDS e uma tabela de *iptables* extremamente restritiva para o Hadoop.

CONCLUSÃO

O ponto de interseção dos casos mencionados no capítulo 4 (locais de implementação e casos de sucesso) é justamente a necessidade de analisar dados que estão cada vez maiores, em vários formatos, em menor tempo, com maior assertividade e que tragam valores diferenciados em um ambiente de competição.

Tudo isso se constitui em desafios para analistas que, não conseguem realizar o trabalho com o mesmo *modus operandi* de tempos atrás, ou seja, por si próprio. Há necessidade de empregar ferramentas adequadas para atingir o objetivo final.

Da mesma forma, decisores não conseguem mais realizar um procedimento e tomada de decisão sem utilizar os dados que acercam o negócio de sua empresa. Qualquer informação que não for passada poderá ter impacto grande. Portanto, trazer a informação de forma rápida e fidedigna pode levar a empresa à liderança ou ao fim, no ambiente de competição.

Cada vez mais, a tecnologia da informação e comunicação evolui. Robôs e procedimentos automatizados têm conseguido dar resultados a demandas de análise de dados, utilizando-se de aprendizado de máquina, inteligência artificial e cognitiva. Profissionais que trabalham com análise poderão ser substituídos por esses instrumentos de automação, caso não acompanhem a evolução. Em outras palavras, conhecer ferramentas de *Big Data* será um requisito para analistas de diversos setores.

O Hadoop é uma forma de levar a capacidade de *Big Data* ao analista. É uma forma de buscar informações em um oceano de dados. Por ser de código aberto, possui grande aceitação no mercado e tem se atualizado no decorrer dos tempos, trazendo resultados às demandas solicitadas.

O software apresentado nesse artigo prima pela simplicidade de implementação. Respondendo às demandas de volumetria, rapidez, diversificação de formatos e veracidade, a ferramenta é capaz de agregar valor em um processo de análise. Portanto, verifica-se que atualmente, ferramentas para análise do *Big*

Data são uma necessidade para o analista de TI, sendo o Hadoop uma opção atrativa e de código aberto.

HADOOP: A NEED FOR ANALYSTS

ABSTRACT: Big Data is a term that is currently in vogue. The need to search a large amount of data in different formats, resulting in fast, assertive and valuable products that opponents do not possess is high in a competitive environment. To do so, analysts need tools that meet these requirements so that their bosses can better carry out the decision-making process. Hadoop is one of these tools and its use will be discussed as the main purpose of this article. Other specific objectives are to characterize Big Data and Hadoop and to show the challenges of implementing the functionality.

Keywords: Big Data, Data, Decision, Tools, Haddop

REFERÊNCIAS

Andrejevic, Mark e Gates, Kelly. Big Data Surveillance: Introduction. Disponível em: https://ojs.library.queensu.ca/index.php/surveillance-and-society/article/download/bds_ed/bds_editorial. Acessado em 15 de setembro de 2019.

Bathal, Gurjit Singh e Singh, Amardeep. Big Data: Hadoop framework vulnerabilities, security issues and attack. Disponível em: https://www.sciencedirect.com/science/article/pii/S2590005619300025. Acessado em 15 de setembro de 2019.

Begoli, E. and Horey, J. (2012). Design principles for effective knowledge discovery from big data. In Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 Joint Working IEEE/IFIP Conference on, pages 215– 218. IEEE

Cetax. Hadoop: o que é, conceito e definição. Disponível em: https://www.cetax.com.br/blog/apache-hadoop/. Acessado em 14 de setembro de 2019.

Chede, Cesar. Software, Open Source, SOA, Innovation, Open Standards, Trends. Disponível em: https://www.ibm.com/developerworks/community/blogs/ctaurion/entry/conhecendo_hadoop?lang=en. Acessado em 10 de setembro de 2019.

Chen, H. (2001), "Knowledge management systems: a text mining perspective", Arizona: Knowledge Computing Corporation.

Cheng, Bin et al. Buiding a Big Data Platform for Smart Cities: Experience and Lessons from Santander. Disponível em: https://www.researchgate.net/publication/280572509_Building_a_Big_Data_Platform_for_Smart_Cities_Experience_and_Lessons_from_Santander_BigData2015-5084. Acessado em 14 de setembro de 2019.

Chin, Ong Eu et al. The Evolution of e-commerce: A case study on Amazon. Disponível em: https://www.academia.edu/30175817/Evolution_of_E-Commerce_Amazon_case_study. Acessado em 14 de setembro de 2019.

Cozza, R., Milanesi, C., Gupta, A., Nguyen, T. H., Lu, C. K., Zimmermann, A., & De La Vernge, H. J. (2011), "Market Share Analysis: Mobile Devices", Gartner Report, Disponível em: http://www.gartner.com/newsroom/id/1689814. Acessado em 08 de agosto de 2019.

Data-Flair. Hadoop Ecosystem and their componentes – A complete tutorial. Disponível em: https://data-flair.training/blogs/hadoop-ecosystem-components/. Acessado em 14 de setembro de 2019.

Dean, Jeffrey e Ghemawat, Sanjay. MapReduce: Simplified Data Processing on Large Clusters. Disponível em: https://static.googleusercontent.com/media/research.google.com/pt-BR//archive/mapreduce-osdi04.pdf. Acessado em 10 de setembro de 2019.

DiFranzo, D., Zhang, Q., Gloria, K., Hendler, J. (2013). "Large Scale Social Network Analysis Using Semantic Web Technologies", AAAI Fall Symposium Series.

Estácio de Sá. Apostila de Tecnologias Avançadas: 2019.

Fayyad, U. et al. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, v. 39, no 11, p. 27-35, 1996.

Ferreira, Pedro Guilherme e Picchetti, Paulo. A era do Big Data e suas implicações para o acompanhamento macroeconômico. Disponível em : https://blogdoibre.fgv.br/posts/era-do-big-data-e-suas-implicacoes-para-o-acompanhamento-macroeconomico. Acessado em 15 de setembro de 2019.

Fiveacts. Afinal, qual a diferença entre Business Intelligence e Big Data. Disponível em: https://www.fiveacts.com.br/afinal-qual-diferenca-entre-business-intelligence-e-big-data/. Acessadoem 15 de setembro de 2019.

Gartner. Big data. Disponível em: https://gartner.com/

it-glossary/big-data/. Acessado em 10 de setembro de 2019.

Gomes, Pedro César Tebaldi. Os 5Vs do Big Data. https://www.datageeks.com.br/5vs-do-big-data/. Acessado em 10 de setembro de 2019.

Hadoop. The Apache Hadoop. Disponível em: https://hadoop.apache.org/. Acessado em 14 de setembro de 2019.

Hazen, Benjamin T. e Jones-Farmer, L. Alisson. Statistical perspective on “big data”. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0925527314001339. Acessado em 10 de setembro de 2019.

Intel. Introdução à Big Data: como avançar com uma implantação bem-sucedida. Disponível em: intel.com.br/content/dam/www/public/lar/br/pt/documents/articles/e7-big-data-planning-guide-webready-por.pdf. Acessado em 10 de setembro de 2019.

Hortonworks. Hortonworks Sandbox with VirtualBox. Disponível em http://hortonworks.com/wp-content/uploads/2015/04/Import_on_VBox_4_07_2015.pdf. Acessado em 15 de setembro de 2019.

Laney, Doug. 2001. 3-D Data Management: Controling Data Volume, Velocity and Variety. Meta Group Research Note, 6 de fevereiro.

Melo, Thiago. Big Data no Judiciário: como a análise de dados pode tornar a justiça mais ágil. Disponível em: https://www.sajdigital.com/tribunal-de-justica/big-data-no-judiciario/. Acessado em 14 de setembro de 2019.

Moreno-Cano, V., Terroso-Saenz, F., and Skarmeta-Gomez, A. F. (2015). Big data for iot services in smart cities. In Internet of Things, 2015 IEEE 2nd World Forum on, pages 418–423. IEEE.

Oliveira, Débora. Hospital norte-americano usa Big Data para analisar informações de pacientes. Disponível em: https://www.itforum365.com.br/hospital-norte-americano-usa-big-data-para-analisar-informacoes-de-pacientes/. Acessado em 14 de setembro de 2019.

Prakash, Om et al. Banking on Big Data: a case study. Disponível em: https://www.researchgate.net/publication/297420730_Banking_on_big_data_A_case_study. Acessado em 14 de setembro de 2019.

Pinto, Helder. Introdução ao Hadoop Yarn. Disponível em: https://www.infoq.com/br/presentations/introducao-ao-hadoop-yarn/. Acessado em 14 de setembro de 2019.

SAS. Hadoop: O que é e qual a sua importância. Disponível em: https://www.sas.com/pt_br/insights/big-data/hadoop.html. Acessado em 14 de setembro de 2019.

Silva, Peterson Ferreira da. A Guerra do futuro já começou e o Brasil enfrenta o desafio do abismo tecnológico. Disponível em: http://ebrevistas.eb.mil.br/index.php/CEEExAE/article/download/2116/1710/. Acessado em 15 de setembro de 2019.

Tan, W., Blake, M. B., Saleh, I., Dustdar, S. (2013), “Social-Network-Sourced Big Data Analytics”. Internet Computing. IEEE Computer Society, v. 17, n. 5, p. 62-69.

Tulloch, Mitch. Business data storage is i setting cheaper – or is it? http://techgenix.com/business-data-storage-costs/. Acessado em 08 de agosto de 2019.

Westcon. Quais os benefícios do Big Data analytics para os negócios. Disponível em: https://blogbrasil.westcon.com/quais-os-beneficios-do-big-data-analytics-para-os-negocios. Acessado em 10 de setembro de 2019.

Eliezer de Souza Batista Junior é bacharel em Ciências Militares pela Academia Militar das Agulhas Negras (AMAN) e Sistemas de Informações (FIP/RO). É pós-graduado de forma latu sensu em Guerra Eletrônica, Cibernética e Análise de Malware; e de forma strictu em Sistema Tático de Enlace de Dados. Atualmente, está realizando pós graduação em Big Data e Análise de dados na faculdade Estácio de Sá e exerce a função chefe da Subdivisão de Gestão de TI do Hospital das Forças Armadas, em Brasília. Pode ser contactado pelo email eliezer.batista@eb.mil.br.

Rôber Yamashita é bacharel em Ciências Militares (Comunicações) pela Academia Militar das Agulhas Negras (AMAN). É pós-graduado em Criptografia e Segurança de Redes pela Universidade Federal Fluminense (UFF). Mestre pela Escola de Aperfeiçoamento de Oficiais (EsAO) e doutorando pela Asia e University (Malásia). Atualmente, é aluno do Curso de Comando e Estado-Maior da Escola de Comando e Estado-Maior do Exército (ECEME) e pode ser contactado pelo email yamashita.rober@eb.mil.br.