Aplicabilidade, utilidade e ganhos do Big Data utilizando a ferramenta Hadoop

Anderson Larcher Franco, Geraldo Magela Almeida Bessa

Centro de Ensino Superior de Juiz de Fora (CESJF)
Caixa Postal 36016-000 – Minas Gerais – MG - Brasil

andersonfranco19@gmail.com, geraldobessa@pucminas.cesjf.br

Abstract. *This article objective is to approach basic Big Data concepts for beginners using software Hadoop as a tool, trying hard to show how it works and the initial concepts, giving at the reader a good knowledge in handling of large volume data. With an increasingly growth of data, technology has been adapting itself in order to keep everything organized and work in more beneficial ways such as aiding users to take decisions or making large systems even more powerful.*

Resumo. *O objetivo desse artigo é fazer um breve estudo sobre Big Data com o uso da ferramenta Hadoop, tentando demonstrar ao máximo como funciona e os conceitos iniciais, dando ao leitor um conhecimento bem sólido em manuseio de dados em grande volume. Com o crescimento de dados no mundo sendo cada vez maior, as novas tecnologias vem se adequando ao mercado, de maneira que tudo possa ser gerenciado de forma organizada e trazer mais benefícios, como ajuda na tomada de decisões dos usuários e fazer com que grandes sistemas possam ser ainda mais poderosos.*

Palavras-chave: Banco de Dados, Big Data, Hadoop.

1. Introdução

É imperceptível aos olhos dos indivíduos que usufruem de sistemas simples ou mais complexos, a quantidade de informação que é disponibilizada para ele de maneira rápida e eficiente. Atualmente, é difícil acreditar que ao fazer pesquisas ou navegar na internet consultando sites de gêneros diversos, as informações que estão chegando até as pessoas, estão sendo processadas de maneira surpreendentemente rápida e de seu absoluto interesse.

É complexo dizer o que pode ser considerado grandioso para uma pessoa, pois cada um tem um conceito e uma ideia diferente quando o assunto é quantidade. Para alguns usuários, 1 gigabyte é considerado uma quantidade exacerbada, como para outros 1000 gigabytes seria algo estimado normal. Com isso, podemos imaginar como é complexo averiguar e manipular os dados de maneira que atenda a concepção de todos.

Com esse desafio, temos o conceito de Big Data, todavia não é tão simples. Ao pesquisar sobre o assunto, encontra-se diversas definições sobre este conceito que embora seja muito utilizado, ainda é desconhecido por muitas organizações. Não apenas por ter inúmeras formas de ser conceituada, mas pelo fato de como foi dito anteriormente, a definição de tamanho de dados pode mudar o entendimento de uma pessoa.

O conceito Big Data é utilizado para caracterizar os dados que excedem a capacidade de processamento de sistemas de banco de dados convencionais. Big Data é muito grande, se move muito rápido, e não se encaixa nas restrições de arquiteturas de banco de dados. Para ganhar o valor a partir desses dados, você deve escolher um caminho alternativo para processá-lo. (Schneider,2012)

O Hadoop é uma plataforma que congrega um grande número de máquinas agrupadas em diversos conjuntos, ou "clusters" ligados a um dispositivo de armazenamento de grande capacidade de processamento (CUTTING, 2015).

Dentre as soluções de Big Data (BigPanda, StreamSets, WebAction, entre outros), o Hadoop foi escolhido para o estudo deste artigo, devido a mesma ser de código aberto e possuir uma popularidade entre as demais ferramentas além de conter grandes contribuidores ajudando no seu crescimento e melhoria. Em seus meios, junto à manipulação de muitos dados, existe a permissão para a criação de um ecossistema de negócios baseados em distribuições particulares, que é basicamente o acesso até mesmo em bases distintas de vários lugares diferentes. Isso diz que ele não deve ser usado por qualquer um, pois embora seja de fácil acesso, exige um breve conhecimento para que se possa navegar na ferramenta e usufruir de todas suas funcionalidades. Com isso, assim como toda plataforma, ele também deve passar por uma avaliação para saber se é válido ou não ser aplicado em uma empresa.

Vendo esses conceitos, se espera que ao final desse artigo se tenha obtido o conceito inicial para poder trabalhar com as tecnologias vistas mostrando de forma simples e sem muitos termos complexos, o ponto de partida para aqueles que desejam iniciar um estudo sobre a ferramenta Hadoop e a manipulação dos dados com o Big Data.

2. Conhecendo Big Data

A primeira coisa a reconhecer segundo Robert D. Schneider (2012), que atualmente atua com planejamento e Construção de bancos de dados de alto desempenho, é que Big Data não tem uma definição única. De fato, é um termo que descreve, pelo menos, três separadas, mas inter-relacionadas tendências que são a captura de lotes de gestão de

informação, trabalhar com muitos novos tipos de dados explorados dessas massas de informação com novos estilos de aplicativos (SCHNEIDER, 2012).

De acordo com Rodrigo Arrigoni, sócio fundador da R18, empresa de comunicação e tecnologia especializada em análise de dados nas redes sociais, o termo Big Data surgiu na década de 1990, porém agora está começando a ganhar seu espaço no mercado de uma forma muito valorizada. Nascido na NASA para descrever grandes conjuntos de dados complexos, o Big Data desafia os limites computacionais tradicionais de captura, processamento, análise e armazenamento informacional. Hoje, na era da informação, temos redes sociais, telefones celulares, GPS e diversos dispositivos móveis que deram o pontapé inicial para o disparo da utilização desse serviço (ARRIGONI, 2013).

É muito comum encontrar definições dessa tecnologia de inúmeras maneiras diferentes. Isso acontece pelo motivo de não ter uma forma de dizer que o tamanho de alguma coisa é grande ou não, e o Big Data permite analisar qualquer tipo de informação em tempo real.

Por ser uma tecnologia que reúne uma quantidade de dados digitais e os cruza posteriormente, o aumento de ganhos com o seu uso é muito grande, tendo um volume e uma velocidade que não era de costume encontrar por volta de dez anos atrás. (ARRIGONI, 2013).

O Big Data veio para modificar a forma como era realizada a manipulação dos dados até os últimos anos, e vem crescendo mais a cada instante. Assim, substitui a centralização, de um Data Warehouse de um seleto conjunto de dados, por uma experiência na locação das informações, e com isso não é mais necessário a junção dos mesmos em um único lugar. Essa tarefa acaba solucionando o problema de lentidão que é gerado pela movimentação de um grande fluxo de informações. Tudo isso é feito com dados não-estruturados, que dependem de um contexto para serem entendido. (PETRY, 2014)

De acordo com cada contexto, começamos a perceber que estamos mais presentes no Big Data do que podemos imaginar. Algoritmos cada vez mais eficientes e complexos fazem sugestões de novos amigos nas redes sociais, tudo sobre uma revisão minuciosa em sua rede de amizades. São também capazes de encontrar padrões eficientes e complexos em acessos de músicas que uma pessoa tenha feito e sugerir algo parecido que o usuário talvez desconheça até então. (PETRY, 2014)

O Big Data não se destina somente às grandes empresas e multinacionais. Pequenas e médias empresas também estão abertas a se beneficiar das vantagens e facilidades que são oferecidas aos seus consumidores. Tal tecnologia que abrange uma área imensurável de empregos sendo utilizada na área de educação, segurança, saúde e não se limitando diretamente ao seguimento lucrativo comercial dos negócios e também na busca por uma quantidade maior de clientes.

2.1. Big Data – Conhecendo os Três V's

O Big Data é conhecido pela sua conceitualização com um termo chamado de três V's, que são volume, variedade e velocidade. Veremos nesse tópico uma definição melhor desses conceitos que ajudam a entender melhor a tecnologia.

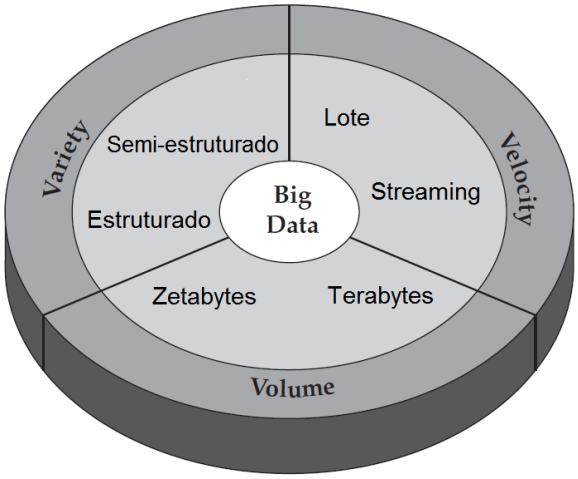

Figura 1 - Divisão dos três termos dentro da tecnologia de Big Data e suas principais características (APANDRE, 2013).

2.1.1. Big Data – Volume

O conceito de volume no Big Data é bem simples. Ele é definição da quantidade de informações digitais que os usuários ou processos de aplicações produzem. Devido o volume de dados ser o ponto de partida do Big Data, esse é o atributo que mais se destaca. Resumidamente os processos de Big Data tem como objetivo encontrar tendências e novos padrões nesses grandes volumes de dados, visto que as ferramentas que estão disponíveis hoje no mercado, não atendem satisfatoriamente quando trabalham com grande volume (PETRY; VILICIC, 2013, p.74-75).

2.1.2. Big Data – Variedade

O Big Data também é caracterizado por sua variedade de dados como arquivos de áudio, fotos, mensagens de celular, comentários e curtidas em redes sociais, histórico de navegações web e ativações de leitora de código de barras. Todos esses tipos de dados criam um grande volume de informações, que na grande maioria das vezes não possui uma forma de estruturação e se tornam informações oriundas de fontes de dados mistas.

Algumas informações geradas a partir da execução de um usuário ou pela própria aplicação, possuem os seus dados criados por um processo automático de um software. Essas informações podem ser alocadas em um Data Warehouse que vem a ser alvo do emprego das aplicações de Big Data além de certa forma possuir uma relação entre si. Isso acaba gerando um novo padrão ou tendência. Além disso, a variedade também é considerada um fator fundamental para a criação do Big Data, visto a dificuldade que ainda se enfrenta nos dias de hoje na tentativa de conseguir abstrair

informações de tipo de arquivos diferentes em que sua essência não possui detalhes semelhantes. Isso acontece por exemplo, ao tentar descobrir um novo padrão de dados acima de uma demasiada combinação de arquivos de imagens com arquivos de áudio (PETRY; VILICIC, 2013, p.74-75).

2.1.3. Big Data – Velocidade

A velocidade no Big Data é tão importante quanto o volume e variedade. Ela adapta a rapidez com as informações criadas, alocadas e selecionadas. Hoje isso é uma necessidade e exigência de todos os envolvidos em um processo digital, como por exemplo nos setores privados e públicos, áreas de telecomunicação, trabalhadores e clientes. Isso se dá devido à importância de conseguir uma resposta em tempo real. Ainda existem muitos softwares que buscam informações em grandes volumes de dados e novos padrões, mas ainda não conseguem processar em tempo real as informações que possuem ciclos menores de execução. Simplesmente por causa de sua demora na manipulação das informações que estão sendo interpretadas (PETRY; VILICIC, 2013, p.74-75).

3. Segurança de Big Data

Segundo Petry e Vilicic (2013), nos dias atuais qualquer usuário que se encontra nessa imensa rede mundial de computadores possui um histórico digital, onde praticamente todas as ações das pessoas estão registradas. Vídeos, fotos, registros particulares entre os mais diversos, estão armazenado digitalmente, o que faz com que cada indivíduo tenha uma forma de perfil digital. Isso tudo pode se encontrar vulnerável e pode expor qualquer um que tenha acesso a esse tipo de conhecimento. Por esse motivo, o cuidado com os dados deve ser extremamente reforçado, visto que as pessoas possuem muito mais dados registrados digitalmente do que na sua própria casa fisicamente. (VILICIC, 2013)

Avaliando uma pesquisa realizada pela universidade de Cambridge, se conseguiu ver resultados surpreendentes na utilização de técnicas Big Data, onde foi relacionado informações contidas no Facebook com o propósito de descobrir algumas informações que ficam omitidas na rede social. O resultado desse estudo foi que em 95% das avaliações feitas, foi possível descobrir dados como a etnia de cada um. Em outros 88% se descobriu o sexo de cada indivíduo, assim como a posição religiosa e política de cada um foi descoberta em 80% dos casos. (PETRY; VILICIC, 2013).

A segurança no Big Data acaba se tornando um caso um pouco delicado por vários motivos. Devido a hoje existir atração por novos conhecimentos, a criação de técnicas para manipular e aperfeiçoar as ferramentas que já existem no Big Data, podem causar vulnerabilidade acentuada, principalmente em software de código aberto e de acesso a todos. Isso acaba abrindo portas para que hackers analisem a ideia principal do projeto, que são oportunidades para estudos e troca de informação para muitos e identificar algumas falhas possíveis que podem resultar em um desprendimento de tempo, que é um fato que gera muitos prejuízos (EMC, 2012).

Devido ao Big Data ser recente, muitas ferramentas de segurança ainda estão em fase de teste e aperfeiçoamento. Em estudos feitos pela Verizon em 2012, onde foi abordado algumas investigações sobre violações de dados, foi mostrado que mais de 90% das invasões atrasam em mais de 24 horas a evolução dos estudos sobre aquele determinado alvo de ataque, isso considerando que cerca de 80% das violações demoram cerca de semanas para serem encontradas. Hoje esses números cresceram ainda mais, pois a quantidade de dados que vem sendo produzido é ainda maior (EMC, 2012).

Um caso que ocorreu em setembro de 2013 que marcou a vulnerabilidade das informações disponíveis na internet, foi o caso da utilização de técnicas de Big Data com ferramentas de espionagem norte-americana a outros países, incluindo até mesmo o Brasil.

Reportagens escritas por jornais de diversos lugares do mundo, citam o norte-americano Edward Snowden, que foi um empregado da empresa Booz Allen que na época prestava seus serviços a Agência Nacional de Segurança dos Estados Unidos (NSA), disponibilizou dados secretos da agência onde foi revelado que o os Estados Unidos praticava atos de espionagem contra vários outros países. Esses documentos foram obtidos pela NSA por meio de programas de computador que conseguiam acessar e-mail, chats online, entre outras inúmeras formas de informações digitais que estavam armazenadas na internet e com a posse dessas informações era possível descobrir formas de ataques terroristas contra os EUA.(EXAME, 2013)

O grande problema nesse caso, foi que todo esse trabalho não se restringiu somente em detectar ataques terroristas, mas também foi descoberto que com todas as informações obtidas pelo software espião, eram coletados dados digitais criados pelas pessoas de todas as partes do mundo, incluindo líderes e chefes de estado. A obtenção desses dados colocou os EUA em uma zona de hegemonia sobre outras nações, pois com tais resultados, os norte-americanos conseguiam analisar os dados traçando perfis de líderes, podendo descobrir a forma que pensavam, de que maneira se comunicavam e até mesmo o que estavam planejando. Com tal conhecimento o controle era muito maior, pois quanto mais se sabia referente às outras nações, mais fácil era o domínio sobre o alvo espionado. (EXAME, 2013)

Todo esse controle e manipulação dos dados só foi possível devido à criação de uma super infraestrutura de captação e processamento de dados, onde são empregadas as técnicas de Big Data. (EXAME, 2013)

Os principais motivos que podem justificar a necessidade da NSA em possuir um imenso poder de captação e processamento de dados é a interceptação indiscriminada de informações que são produzidas pelos usuários convencionais da internet e o fato de grande parte das informações que circulam na rede estarem criptografadas. Coisas que vão de dados de transações financeiras até mensagens eletrônicas, circulam criptografadas pela internet. Por esse motivo existe a necessidade de computadores mais potentes para análise e processamento de informações que são capazes de quebrar a criptografia e descobrir o significado de cada dado que está sendo analisado.(EXAME, 2013)

4. Big Data - Desafios

Neste tópico serão mostrados grandes desafios que a era do Big Data tende a enfrentar. Em alguns casos, será visto que já é comum tal tipo de dificuldade além de ser possível contorná-la.

4.1. Big Data – Direito do Consumidor

No Big Data, o direito do consumidor é uma chave importante na manipulação de informações devido à facilidade que se tem hoje em dia na aquisição de produtos digitais e na produção, de certa forma incontrolável, de informações de redes sociais, e-mail, vídeos e fotos feitos pelos usuários, principalmente via celular. Isso torna necessário a criação e a regulamentação de leis que possam defender os direitos de propriedade e privacidade das informações particulares de cada cliente. Essa medida tem como ideia, a preservação e resguardo de possíveis crimes como por exemplo, roubo ou utilização de informações privadas de forma indevida. Uma outra medida a ser tomada por parte do poder público, é fazer um incentivo correto e, totalmente, direcionado às empresas, para que façam o uso dos dados a favor da sociedade e não usa-los de maneira prejudicial. (HUGO, 2014)

4.2. Big Data – Competitividade Comercial

No Big Data, a competitividade comercial vem trazendo o consumidor para dentro da tecnologia de maneira que o marketing consiga atrair sua atenção. Com tamanha grandeza que a tecnologia possui, organizações de grande porte vão conseguir produzir aplicações cada vez mais completas e complexas, acessíveis e de fácil manuseio, que fará com que a sociedade se sinta atraída pela busca desses benefícios oferecidos. Entretanto, essa oferta de produtos de certa forma tão surpreendentes cria uma forma de cenário de competitividade muito desnivelado, onde as pequenas empresas acabam se tornando inevitavelmente alvo de uma soberania por parte das gigantes que possuem mais recursos, prejudicando por outro lado a competitividade comercial e principalmente o lado do cliente.

Grandes organizações que são do setor privado em sua grande maioria devem ser incentivadas a trabalharem em uma forma de produção de software com código aberto, visando que esses programas Open Source permitam a construção de novas ideias e o aperfeiçoamento das tecnologias, diminuido assim um pouco da competitividade e fazendo que se tenham ganhos no conhecimento. Além disso, através de ferramentas adequadas, a análise preditiva do comportamento do consumidor oferece a possibilidade de determinar como o público reagirá às estratégias de marketing, com base em padrões anteriores de comportamento. (Hekima, 2014)

4.3 Big Data – Mão de Obra Especializada

Segundo a Datastorm (empresa especialista em Big Data Analytics), profissionais para lidar com estruturação de dados ainda são “escassos” no país (também são poucas empresas especializadas que fornecem soluções concretas nessa área). Tudo isso faz com que empresas brasileiras apresentem certa resistência na implantação de algum

projeto que use dados como fonte primária de valor. De fato, montar uma equipe interna especializada pode ser um pouco caro (e difícil), principalmente pela falta de profissionais qualificados. Por isso, a contratação de empresas especializadas é algo a se pensar para extrair valor dos dados. (DATASTORM, 2016)

Há ainda alguns profissionais e empresas mais descrentes que relataram a falta de pesquisas ou provas da vantagem do Big Data nos negócios como motivo para não se aventurarem nessa área. (DATASTORM, 2016)

5. Hadoop e seus benefícios

Segundo Doug Cutting, arquiteto-chefe que ajudou a criar Apache Hadoop em um caso de necessidade, os dados da web explodiram e cresceram muito além da capacidade dos sistemas tradicionais de lidar com isso. Ainda assim segundo ele, o Hadoop foi inicialmente inspirado por trabalhos publicados pelo Google descrevendo sua abordagem para lidar com uma avalanche de dados, e desde então se tornou o padrão de fato para o armazenamento, processamento e análise de uma grande quantidade de dados. (ITFORUM365 , 2015)

Tais tecnologias como o Hadoop da Apache, são de extrema importância para ajudar as empresas a gerirem grandes quantidades de dados. Grandes organizações como a Netflix, Twitter e NASA, hoje usam a tecnologia para poder alavancar seus ganhos com análise de dados em todas as partes. (COMPUTERWORLD, 2015)

Vivendo o conceito de Big Data, a plataforma aberta de computação distribuída ganhou um grande impulso como forma de um dispositivo para poder lidar com a tecnologia, onde as empresas procuram extrair grandes valores de fluxos de dados em seus sistemas de informação.(COMPUTERWORLD, 2015)

Outro ponto que a Cloudera afirma é que, o Hadoop pode lidar com todos os tipos de dados de sistemas distintos: estruturados, não-estruturados, arquivos, imagens, arquivos de áudio, registos de comunicações de e-mail, log e tudo sobre qualquer coisa que você pode pensar, independentemente do seu formato nativo. Mesmo quando diferentes tipos de dados forem armazenados em sistemas alheios, você pode despejar tudo em seu cluster Hadoop, sem a necessidade prévia de um esquema. Em outras palavras, você não precisa saber como você pretende consultar seus dados antes de armazená-lo. O Hadoop permite que você decida ao longo do tempo, que pode revelar perguntas que você nunca pensou em imaginar. (CLOUDERA, 2015)

Tornando todos os seus dados utilizáveis e não apenas o que está em seus bancos de dados, o Hadoop permite que você veja as relações que estavam escondidas antes de revelar respostas que sempre estiveram fora de alcance. Você pode começar a fazer mais decisões baseadas em dados concretos em vez de palpites e olhar para conjuntos de dados completos e não apenas amostras. (CLOUDERA, 2015)

Ao se ver como são feitas as plataformas tradicionais segundo a CIO da Apache, sabe-se que o Hadoop é capaz de armazenar qualquer tipo de dado no seu formato

nativo e realizar em escala, uma variedade de análises e transformações sobre esses dados. (ITFORUM365 , 2015)

Isso já ajuda em grande parte do armazenamento onde não é preciso ter uma conversão dos dados para um formato especifico.

Com isso, a própria Cloudera afirma a existência de algumas vantagens ao escolher o Hadoop entre outras tecnologias tradicionais, tendo pontos fortes exclusivos, como:

- Um armazenamento em grandes volumes de dados seja qual for tipo de dado (ex.: escala de petabytes).

- Desempenho em análises de dados complexas para dados diversos como web logs, e-mail, registros de detalhes de chamadas, dados sociais, mensagens de texto, dados de dispositivos instalados em máquinas, dados de rede, vídeos e imagens.

- Hadoop oferece uma estrutura de alto desempenho capaz de processar grandes volumes de dados, à medida que as demandas organizacionais mudam.

- As organizações que rodam o Hadoop – de grandes segmentos e sistemas menores, na escala de terabytes – caracterizam a capacidade dele escalar tanto para cima quanto para baixo como uma vantagem definitiva no que se diz respeito ao desempenho e relação entre eficiência e custo.

- Há uma vantagem de custo significativa com o uso do Hadoop, já que é um software de código aberto que roda em hardware comum.

Vendo que exitem alguns ganhos segundo a CIO da Apache, ao mesmo tempo há vários desafios associados a sua implementação. Podem haver dificuldades para encontrar recursos pela falta de desenvolvedores capacitados e especialistas de dados habilitados aos tipos de tarefas e projetos selecionados para o Hadoop. As tarefas de dados desempenhadas precisam ser integradas com o restante dos sistemas de TI. Adicionar scripts a estas tarefas pode muitas vezes causar problemas quando as organizações querem mover dados para dentro e para fora do software com consistência, confiabilidade e eficiência. Outros desafios incluem a gestão do fluxo de processamento e governança de dados. Implementar o Hadoop de modo que supere estes desafios é importante para maximizar os ganhos. (ITFORUM365, 2015)

Outro conceito que vem sendo muito usado é o HDInsight. Ele é uma implementação de nuvem da Microsoft Azure na pilha de tecnologia Apache Hadoop, que está se expandindo rapidamente e também pode ser uma opção para a análise de Big Data. Segundo a Microsoft, exitem vantagens em usar o Hadoop na nuvem. Como parte do ecossistema de nuvem do Azure, o Hadoop no HDInsight oferece uma série de benefícios (Microsoft Azure, 2015), entre eles:

- O provisionamento automático de clusters Hadoop. É muito mais fácil criar clusters HDInsight do que configurar clusters Hadoop manualmente.

- Componentes do Hadoop de última geração.

- Alta disponibilidade e confiabilidade dos clusters.

- Armazenamento de dados eficiente e econômico com o armazenamento de Blob do Azure, uma opção compatível com o Hadoop.

- Integração com outros serviços do Azure, incluindo aplicativos Web e Banco de Dados SQL.

- Baixo custo de entrada.

6. Hadoop: Pequenas Empresas

Analisando as pequenas empresas, segundo Thoran Rodrigues (CEO da Big Data Corp), o mais interessante de todo esse fenômeno para as startups e pequenas empresas é um nivelamento dentro do jogo dos negócios. Thoran ainda afirma que antes de termos essa abundância de informações, somente quem podia ter conhecimento maior em relação ao consumidor eram as grandes empresas. (TERRA, 2014)

O Hadoop permite extrair um grande valor na organização, tratamento e consumo de dados com um baixo custo e uma reivindicação até então que não havia sido pensada anteriormente, pelo menos no que diz respeito ao orçamento podemos dizer que a sua utilização está ao alcance até mesmo das pequenas empresas. (POWERDATA, 2015)

Além disso, existem não só pequenas mas como também médias empresas interessadas no Hadoop considerando a crescente importância que está se tornando a análise de Big Data para a tomada de decisões estratégicas, redução de custos e melhoria dos produtos e serviços no ambiente de mercado atual. (POWERDATA, 2015)

Segundo a Powerdata (site de pesquisas relacionadas a crescimentos tecnológicos), embora seja verdade que em comparação com as grandes organizações, as pequenas empresas terão de enfrentar dificuldades adicionais para a sua falta de recursos, essas afirmações acabam por não ser definitivas. Independentemente do tamanho da empresa a chave para decidir a usar Hadoop depende acima de tudo, da necessidade de gerenciar grandes volumes de dados mas como também poder ter tudo o que a tecnologia oferece para ser usado posteriormente. (POWERDATA, 2015)

Outra afirmartiva feita pela Powerdata é o Hadoop ser uma solução para grande volume de dados que não podem ser armazenados e analisados juntamente de infraestruturas tradicionais devido ao volume e tipo de Big Data, contendo informações provenientes de uma variedade de fontes, tais como vídeos, arquivos de imagem, dados de sensores e máquinas, dados transacionais e interações, redes sociais, hábitos e metadados. (POWERDATA, 2015)

6.1 Hadoop: Solução de Baixo Custo para as Empresas Extraírem Valor de Big Data

Ao se analizar os gastos para extrair valor de Big Data, o Hadoop é um sistema de baixo custo que armazena informações heterogêneas, que é o contraditório dos bancos de dados relacionais, onde são usados servidores individuais. Em contrapartida, o Hadoop usa muitos nós, multiplicando conforme necessário os fluxos de dados o que também é mais barato. Dessa maneira é possível ver que o volume de dados é enorme e também cresce exponencialmente assim como um universo de dados maior. (Schneider,2012)

Ainda assim, pequenas e médias empresas estão confiantes em mergulhar nessa ideia de absolver o Big Data, mesmo visto que muitas ainda sentem medo quando se tem uma abordagem de novas tecnologias. Mesmo que aconteça isso, como concluiu um estudo recente, a verdade é que não só pequenas e médias empresas, mas o mundo dos negócios como um todo, utiliza o Hadoop na produção, isso devido ao receio de se ter grandes gastos para pequenas soluções. (Schneider,2012)

No entanto, ter o crescente fluxo de dados de frente faz com que tenhamos a tecnologia que permite extrair valor delas. Logo, o Hadoop é uma grande oportunidade para todas aquelas empresas que procuram no ambiente competitivo de hoje, deixar seus receios para trás é fazer o que os outros não fazem, construindo algo melhor para ter um sucesso crescente a frente daquele que não os implementam.(Schneider,2012)

6.2 Todos os Dados são Iguais

No passado o armazenamento de dados costumava demandar grandes investimentos. Já nos últimos cinco anos, pequenas, médias e grandes empresas obtiveram a descoberta que tinham que preservar e manter a quantidade de um conjunto de dados muito grande, como e-mail, resultados de pesquisa, estoque, venda, clientes, entre outros. Porém, tentar lidar com tudo isso baseando-se em um sistema de gerenciamento de banco de dados relacional seria uma proposta onerosa. (COMPUTERWORLD, 2015)

Atualmente com a chegada de eventos como os citados acima, organizações que arriscavam manter o gerenciamento de dados em dia a um custo acessível, tiveram que passar a colher amostras para criar subconjuntos de dados menores. Logo, essa pequena amostra de dados é automaticamente classificada como acordo de suposições. Por exemplo, as prioridades dos dados no comércio eletrônico podem ser baseadas em suposições de que os dados de cartão de crédito possuem mais importância do que o produto vendido e que por sua vez, pode ser mais importante ainda do que o retorno de cliques (click-through). (COMPUTERWORLD, 2015).

Se por ventura a idéia é desenvolver um modelo de negócios baseado em um conjunto de pressupostos, seria difícil extrair informações para tomar decisões. Porém se a informação fornecida for baseada nesses pressupostos, o que aconteceria se eles estivessem errados? A resposta é simples. Como foi reduzida a amostra de dados, qualquer outro cenário de um novo negócio teria de usar esses mesmos conjuntos de

dados e com isso, os dados originais seriam perdidos para sempre. Sem contar que por causa do alto custo de um sistema de armazenamento baseado em RDBMS, esses dados ficariam isolados na organização (COMPUTERWORLD, 2015).

Já o setor de Vendas, o Marketing e assim por diante, teriam seus próprios dados. Assim as decisões são limitadas a cada parte da organização e não a todas. Com o Hadoop, não se realiza evidências porque todos os dados conversam entre si e esse talvez seja o maior benefício do Hadoop, mas ainda assim muitas vezes, permanece escondido atrás da ideia de redução dos custos da tecnologia. A diminuição de amostras obriga adivinhar que parte dos dados será maior e mais importante do que o resto", segundo Murthy. Ele ainda completa que no Hadoop todos os dados têm o mesmo valor (COMPUTERWORLD, 2015).

Logo, já que todos os dados são iguais e também estão disponíveis a qualquer momento, a companhia pode desenvolver cenários de negócios diferentes, não tendo limitação e sempre utilizando dados originais. Além disso, os dados que foram previamente isolados agora podem ser acessados e compartilhados para analisar as atividades da organização de forma mais global. Mas existe uma diferença na percepção de dados, que é enorme. Uma vez que os dados são armazenados da forma que são, é possível reduzir custos operacionais na gestão de informações associadas com as atividades de transformar e carregar operações.(FENAINFO, 2015).

Não se pode esquecer, no entanto, o benefício mais comentado do Hadoop: a redução de custo. Isso porque, o framework é baseado em código aberto sob a licença Apache Software sem custos de licenciamento para a base de software.(FENAINFO, 2015).

6.3 Hadoop: Usar ou não?

Apesar dos benefícios potenciais da implementação do Hadoop, ainda existem algumas limitações que a organização deve ter em mente antes de saltar para esse universo. Primeiramente, caso a empresa gere relatórios interativos secundários a partir de seus dados ou os utiliza em operações complexas em várias etapas, uma solução RDBMS ainda pode ser a melhor aposta visto que o Hadoop não é particularmente forte nessas áreas. Se os dados da organização são atualizados e alteradas por meio de inserções e eliminação, essa é outra razão para não apostar em Hadoop. (COMPUTERWORLD, 2015).

A Cloudera, (fornecedora comercial do Hadoop) tem como funcionário Doug Cutting, que é um dos inventores do framework e utiliza um modelo de núcleo aberto. A base de software Hadoop é livre, mas extensões Cloudera estão sujeitas a um licenciamento. A Hortonworks, que Murthy fundou com outros membros da equipe de Hadoop para o Yahoo! no início de 2011, mantém todo o software livre e de código aberto e gera receitas por meio de seus programas de treinamento e suporte. O Hadoop possibilita economia, já que não necessita de um hardware caro nem de um processador de alta potência. Qualquer servidor convencional ligado à rede do Hadoop funciona corretamente. Isso significa que um nó do Hadoop só precisa de um processador, um cartão e algumas unidades de disco rígido, com um custo total de cerca de US$3.000,00

(três mil dólares), enquanto um sistema RDBMS pode custar entre €8.000,00 (oito mil euros) e €11.000,00 (onze mil euros) por terabyte. Essa diferença substancial faz com que o Hadoop esteja dentro das empresas. (FENAINFO, 2015).

No entanto, é preciso tomar cuidado para que todo o investimento não comprometa o plano de migração para Hadoop. Outro ponto a ser considerado é o conhecimento técnico necessário para lidar com esse novo mundo. De acordo com analistas do mercado, a demanda por pessoal qualificado pode aumentar os custos do projeto. Nos Estados Unidos, por exemplo, a disputa por engenheiros qualificados em Hadoop tem sido tão acirrada, que dois dos maiores atores da plataforma [Google e Facebook] entraram em uma guerra de lances para atrair engenheiros (COMPUTERWORLD, 2015).

Independente do software que a organização implemente, ela deve estar preparada para investir pesado na equipe de Hadoop. Dependendo das necessidades e localização, a companhia poderia ter de investir entre 100 mil dólares e 150 mil dólares por ano. Porem, apesar de ter de pagar um extra para o administrador de Hadoop, os benefícios da tecnologia atraem cada vez mais companhias que decidem obter reduções significativas de custo no longo prazo (COMPUTERWORLD, 2015).

6.4 Hadoop: Mercado

O Hadoop vem crescendo no mercado, porém ainda é notável também a falta de mão de obra qualificada para se trabalhar com a ferramenta. Mesmo assim, o mercado Hadoop está se expandindo rapidamente e com isso encontramos um mercado promissor, com grandes perspectivas de crescimento para grandes empresas e para os postos de trabalho (CETAX, 2015).

De acordo com o site Cetax, a pesquisa “The Big Data Executive Survey” feita pela NewVantage Paterns, foi realizada com 90 executivos suportando as 1000 maiores empresas dos EUA, 90% das organizações pesquisadas já estão fazendo algo com Big Data.

Segundo a Cetax, existe uma previsão para que até o fim de 2015, a demanda Big Data gerará 4,4 milhões de postos de trabalho na indústria de TI em todo o mundo aproximadamente.

Só nos EUA 1,9 milhões de postos de trabalho serão criados, afetando diretamente a indústria de TI. Logo, a projeção é de uma figura gigantesca de 6 milhões de empregos nos Estados Unidos, criada por esta nova economia da informação nos próximos 4 anos, mas não há talento suficiente (CETAX, 2015).

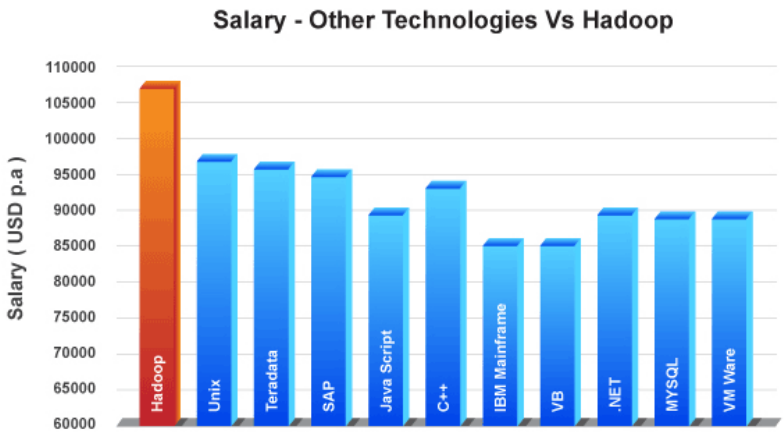

Figura 2 - Gráfico comparativo do salário de profissionais Hadoop com outras tecnologias (CETAX, 2015)

7 Conclusão

Pode-se concluir com o conhecimento repassado neste artigo que o Big Data é a tecnologia que tende a estar presente nos grandes cenários de dados atualmente. Sua grandeza faz com que a organização e a importância das informações possam ser aproveitadas da maneira que for necessária, fazendo com que quem usufrui dessa tecnologia, possa aproveitar mais do que lhe é oferecido.

Esse artigo tem como contribuição, um conhecimento objetivo e simples sobre a tecnologia do Big Data junto a conceitualização da ferramenta Hadoop para profissionais que estão começando seus estudos na área de dados e ainda não possuem um conhecimento básico para começar sua jornada na era das informações. Com os exemplos passados e os conceitos demonstrados no decorrer da leitura do artigo, é possível ter uma decisão de partida de onde começar a trabalhar. Não é feita uma comparação entre outras ferramentas pois a ideia não é apoiar ou incentivar o uso de qualquer software e sim demonstrar do que o Hadoop pode ser capaz de trazer para seus usuários, principalmente aqueles que precisam manipular grandes massas de dados e extrair informação das mesmas.

Tais comparações de ferramentas do mesmo porte, exigem um conhecimento amplo da área de Big Data. The NOSQL e Apache Spark são exemplos de outras grandes ferramentas capazes de manipular dados em grandes quantidades. Ambas ferramentas, incluindo o Hadoop, exigem um conhecimento básico de análise e manipulação de dados para seu manuseio de forma correta.

Quanto ao Hadoop, conclui-se que ele pode ser a ferramenta mais popular devido a alguns fatores presente hoje no cenário de Big Data, como a uma

facilidade maior de obtenção de material didático para aprendizado e também graças a outras circunstancias, como já dito antes, possuir grandes investidores e ser de código aberto. Sua grandiosidade está presente em empresas por todas as partes, sendo algumas delas multinacionais e empresas de grande porte. Isso torna ele uma ferramenta mais visada devido à facilidade de acesso e obtenção de treinamentos.

Assim como qualquer outra ferramenta, o Hadoop possui suas desvantagens em relação a outras como no gerenciamento do aglomerado, que é visto como um dos maiores problemas sofridos pelos desenvolvedores de computação paralela que ocorre no momento de depurar a execução da aplicação e de analisar os logs que estão distribuídos. Infelizmente, com o Hadoop, também ocorrem essas mesmas dificuldades, entretanto, existem subprojetos do ecossistema Hadoop que procuram minimizar esse problema. Outra desvantagem que podemos concluir é que o Hadoop possui um único nó mestre. A arquitetura do Hadoop tradicionalmente utiliza várias máquinas para o processamento e armazenamento dos dados, porém, contém apenas um nó mestre. Esse modelo pode se tornar restritivo a medida que essa centralidade causar empecilhos na escalabilidade e ainda criar um ponto crítico, suscetível a falha, uma vez que o nó mestre é vital para o funcionamento da aplicação, e devido a isso a decisão de qual ferramenta usar, vem por parte do usuário. (Alfredo Goldman, 2016)

Outro ponto a ser levado em consideração durante o estudo desse artigo, é que no ponto de vista de quem trabalha com grandes volumes de dados ou está apenas começando, talvez trabalhar com uma ferramenta open source seja uma opção favorável devido a custos menores. Mas existem aqueles que preferem trazer para dentro de suas empresas um pouco de estabilidade ao arriscar trabalhar com uma ferramenta de código aberto.

8 Referências

CAVALCANTE, Vitor. Pai do Hadoop fala sobre BD transacional, concorrência e big data nas empresas. http://itforum365.com.br/noticias/detalhe/116842/pai-do-hadoop-fala-sobre-bd-transacional-concorrencia-e-big-data-nas-empresas. 2015.

CETAX, Porque estudar o Hadoop. http://www.cetax.com.br/porque-estudar-hadoop. 2015.

COMPUTERWORLD. Hadoop cimenta importância para Big Data. http://www.computerworld.com.pt/2012/06/19/hadoop-cimenta-importancia-para-big-data/. 2012.

COMPUTERWORLD. Hadoop ou sistemas de gestão de base de dados relaciona. http://computerworld.com.br/tecnologia/2012/03/16/hadoop-ou-sistemas-de-gestao-de-base-de-dados-relacional. 2012.

COMPUTERWORLD. Cinco coisas que você precisa saber sobre Hadoop e Apache Spark. http://computerworld.com.br/cinco-coisas-que-voce-precisa-saber-sobre-hadoop-e-apache-spark. 2015.

COMPUTERWORLD. Veja como as empresas usam Hadoop para reforçar seus projetos de big data. http://computerworld.com.br/como-sete-empresas-usam-hadoop-para-reforcar-aplicacoes-de-big-data. 2016.

CLOUDERA. Apache Hadoop. http://www.cloudera.com/content/www/en-us/products/apache-hadoop.html. 2016.

CURIOSO, Luís. Simplifique a integração de Big Data no Hadoop. http://cio.com.br/opiniao/2013/03/04/simplifique-a-integracao-de-big-data-no-hadoop/. 2013.

DATASTORM. Big Data no Brasil: a era da informação já começou. http://datastorm.com.br/big-data-no-brasil-a-era-da-informacao-ja-comecou. 2016.

EMC. A realidade sobre gerenciamento de segurança e big data. White Paper. https://brazil.emc.com/collateral/white-papers/h0812-getting-real-secuirty-management-big-data-wp.pdf. 2012.

FENAINFO, Hadoop ou sistemas de gestão de base de dados relacional. http://fenainfo.org.br/info_ler.php?id=40597. 2013.

GOLDMAN, Alfredo. Capítulo 3 Apache Hadoop: conceitos eóricos e práticos, evolução e novas possibilidades. http://www.ime.usp.br/~ipolato/JAI2012-Hadoop.pdf. 2012.

GREGO, Mauricio. Como a NSA consegue espionar milhões de pessoas nos EUA. http://exame.abril.com.br/tecnologia/noticias/como-a-nsa-consegue-espionar-milhoes-de-pessoas-nos-eua. 2013.

GONÇALVES, Victor Hugo. Big data, Propriedade Intelectual e a segurança de informação: o cidadão comum na nuvem de dados. http://www.migalhas.com.br/dePeso/16,MI201268,61044-Big+data+Propriedade+Intelectual+e+a+seguranca+de+informacao+o. 2014.

GRONLUND, C.J. Introdução ao Hadoop no HDInsight: análise de Big Data e processamento na nuvem. https://azure.microsoft.com/pt-br/documentation/articles/hdinsight-hadoop-introduction/. 2016.

HEKIMA, Como o Big Data está revolucionando as estratégias das empresas. http://www.bigdatabusiness.com.br/como-o-big-data-esta-revolucionando-as-estrategias-das-empresas/. 2014.

PANDRE, Andrei. Datawatch has 3 Vs, dows Visualisation now! https://apandre.wordpress.com/2013/11/19/datawatch/. 2013.

PETRY, A. Vida Digital: O Berço do Big Data. Revista Veja, São Paulo, Maio. p.71-81, 2013.

PETRY, André; VILICIC, Filipe. A era do Big Data e dos algoritmos está mudando o mundo. Revista Veja. n.20, 2013. p.70-81.

POWERDATA, Big Data: ¿Hadoop es sólo para las grandes empresas?. http://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/402575/Big-Data-Hadoop-es-s-lo-para-las-grandes-empresas. 2015.

RODRIGUES, Thoran. Big data ajuda pequena empresa a entrar na mente de clientes. http://economia.terra.com.br/vida-de-empresario/big-data-ajuda-pequena-empresa-a-entrar-na-mente-de-clientes,7171a403d01c8410VgnVCM4000009bcceb0aRCRD.html. 2014.

ROGENSKI, Renato. Uma entrevista didática sobre Big Data. http://exame.abril.com.br/tecnologia/noticias/uma-entrevista-didatica-sobre-big-data. 2013.

SALES, Robson. O que é Big Data? Conheça essa tecnologia de monitoramento. http://www.techtudo.com.br/artigos/noticia/2012/04/voce-sabe-o-que-e-big-data-tecnologia-que-pode-monitorar-sua-vida-ja-movimenta-us-70-bi-no-mundo.html. 2012.

SCHNEIDER, R. D. Hadoop For Dummies, Special Edition. Mississauga, CAN: John Wiley & Sons Canada, 2012. 41 p.

TARIFA, Alexandre. O que é big data e como usar na sua pequena empresa. http://exame.abril.com.br/pme/noticias/o-que-e-big-data-e-como-usar-na-sua-pequena-empresa. 2013.