Faculdade de Tecnologia de Americana "Ministro Ralph Biasi"

Curso Superior de Tecnologia em Segurança da Informação

João Victor Maraia Franco

Murilo Henrique Gomes

COMO O USO DO BIG DATA PODE INFLUENCIAR NA TOMADA DE DECISÃO

AMERICANA – SP

2020

Faculdade de Tecnologia de Americana "Ministro Ralph Biasi"

Curso Superior de Tecnologia em Segurança da Informação

João Victor Maraia Franco

Murilo Henrique Gomes

COMO O USO DO BIG DATA PODE INFLUENCIAR NA TOMADA DE DECISÃO

Trabalho de Conclusão de Curso desenvolvido em cumprimento à exigência curricular do Curso Superior de Tecnologia em Segurança da Informação

Orientador: Profª. Esp. Juliane Borsato Beckedorff Pinto

AMERICANA – SP

2020

FICHA CATALOGRÁFICA – Biblioteca Fatec Americana - CEETEPS
Dados Internacionais de Catalogação-na-fonte

F895c FRANCO, João Victor Maraia

Como o uso do Big Data pode influenciar na tomada de decisão. / João Victor Maraia Franco, Murilo Henrique Gomes. – Americana, 2020.

46f.

Monografia (Curso Superior de Tecnologia em Segurança da Informação) - - Faculdade de Tecnologia de Americana – Centro Estadual de Educação Tecnológica Paula Souza

Orientador: Profa. Esp. Juliane Borsato Beckedorf Pinto

1 Redes de computadores I. PINTO, Juliane Borsato Beckedorf II. Centro Estadual de Educação Tecnológica Paula Souza – Faculdade de Tecnologia de Americana

CDU

João Victor Maraia Franco
Murilo Henrique Gomes

Trabalho de Conclusão de Curso desenvolvido em cumprimento à exigência curricular do Curso Superior de Tecnologia em Segurança da Informação sob a orientação da Professora Especialista Juliane Borsato Beckedorff Pinto

Americana, 12 de Dezembro de 2020

BANCA EXAMINADORA

_____

Profa. Ana Lucia Spigolon

_____

Prof. Maxwell Vitorino da Silva

AGRADECIMENTO

À nossa orientadora Professora Especialista Juliane Borsato Beckedorff Pinto, amigos e familiares que nos apoiaram e acompanharam nossa caminhada até a entrega do trabalho de graduação.

RESUMO

Esta monografia apresenta um estudo sobre o Big Data e como a *Cambridge Analytica* pode ter utilizado o Big Data a seu favor. Dados são uma das principais ferramentas utilizadas na tomada de decisões, sendo algo muito importante e valioso dentro de uma organização. Eles podem auxiliar a empresa a definir novos investimentos, previsões, podendo ajudar na ciência, colaborando em descobertas valiosas, corte de gastos, otimizar processos, entre outras operações. Algumas empresas enxergaram uma oportunidade de utilizarem o Big Data para gerenciar grandes quantidades de dados no gerenciamento de métrica, porém, essas informações podem ser usadas para alterar resultados, como campanhas políticas e a fim de criar métricas e gerar algoritmos de personalidade dos usuários de redes sociais. Este trabalho usará de exemplo o caso da *Cambridge Analytica, que* conseguiu gerar uma ampla visão do que tinha que ser feito nas eleições presidenciais de 2016 nos EUA. E para relatar como tudo isso aconteceu é preciso entender o conceito de como o Big Data funciona e como é possível gerar pequenas leituras utilizando algumas ferramentas.

Palavras-Chave: Cambridge Analytica. Big Data. Dados. Política. Métricas.

ABSTRACT

This monograph presents the study on Big Data - as Cambridge Analytica availed big data to your advantage. Data is one of the main tools used in decision making, being something crucial and valuable within a corporation. They can help the company to define new investments, forecasts, being able to help in science collaborating on valuable discoveries, cutting expenses, optimizing processes, among other operations. Some companies saw an opportunity to use Big Data to manage large amounts of data without managing metrics and avoiding losses, however, this information being used to change campaign results and to create metrics and generate personality algorithms for Cambridge network users was able to generate a wide view of what had to be done in the US presidential elections in 2016. And to describe about how it all happened, it is necessary to understand the concept of how Big Data works and how it is possible to generate small readings using some tools.

Keywords: Cambridge Analytica. Big Data. Data. Politics. Metrics.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

CA - Cambridge Analytica

AWS - Amazon Web Service

GCP - Google Plataforme Cloud

SUMÁRIO

1. INTRODUÇÃO

O Termo Big Data foi usado pela primeira vez em 1997 para nomear a enorme quantidade de dados gerados diariamente na Internet. Todos os dias, são criados 2,5 quintilhões de ‘bytes’ de dados — tanto que 90% dos dados no mundo hoje foram criados nos últimos dois anos (OLIVEIRA, 2015).

Hoje os dados são essenciais para evolução da ciência e também para gerenciar métricas e prevenir perdas. O Big Data ficou famoso pela sua diversificação de dados; quando se fala de Big Data, está se falando sobre armazenamento e processamento de qualquer dado, seja ele de qualquer formato.

Além disso, com o crescimento rápido da informação nos últimos anos, ficou muito mais fácil realizar coletas de dados, seja ela realizada em navegação *Web,* mídias sociais, dados transacionais de diferente natureza.

Em julho de 2017, o *Facebook* alcançou cerca de 2 bilhões de usuários ativos (G1, 2017). Em outras palavras, o *Facebook* tem uma média de ⅓ da população mundial acessando diariamente. Para alguns esse alcance tem um enorme poder de engajamento com pessoas compartilhando informação e dando opinião na rede, sendo assim uma fonte muito valiosa. Com o enorme crescimento dos dados gerados todos os dias através de mídias sociais, navegação *Web*, *post* no *Facebook, Twitter, Instagram*, publicações compartilhadas, tem chamando a atenção de grandes companhias para realizar métricas e aprendizado de máquinas.

Um exemplo disso é a empresa *Cambridge Analytica* que usou um algoritmo para salvar e processar os dados através de um aplicativo dentro da plataforma *Facebook* de modo a criar métricas e gerar algoritmos de personalidade dos usuários da rede social, conseguindo gerar assim uma ampla visão do que tinha que ser feito nas eleições presidenciais de 2016 nos EUA.

Esta monografia tem como objetivo, fazer uma introdução sobre o conteúdo do Big Data que serão demonstrados alguns conceitos e práticas utilizando algumas ferramentas, e como a *Cambridge Analytica* utilizou o Big Data para armazenar informações e realizar o mapeamento de comportamento para analisar e ver quais eleitores podiam ou não mudar de ideia em relação às eleições.

No capítulo dois, será abordado uma pequena introdução de quem é a *Cambridge Analytica*, conceito do Big Data e o crescimento dos dados na *internet.*

No capítulo três, será abordado grande parte do processo teórico de como a

empresa *Cambridge* utilizou tecnologia e ferramentas de alto nível de complexibilidade para criar um grande sistema de mineração e processamento de dados. Nesse mesmo capítulo, será apresentado em forma teórica sobre *fake news*, manipulação de dados e engenharia social.

No capítulo quatro, serão mostrados possíveis cenários para o Big Data e será abordado de forma teórica algumas ferramentas *Open Source* (Código Aberto), e também, como é possível realizar a leitura dos dados de forma prática utilizando uma das ferramentas de Big Data.

Por fim, no quinto capítulo, serão comentadas algumas considerações baseadas no resultado da leitura e processamento de dados usando algumas ferramentas de Big Data e algumas sugestões que poderiam ser utilizadas para trabalhos futuros.

2. REVISÃO BIBLIOGRÁFICA

Hoje em dia é impossível citar a internet e não referenciar o *Google* ou o *Facebook.* Com o aumento considerável das tecnologias e os meios de comunicação, várias pessoas procuram os meios mais fáceis para se comunicarem uns com os outros ou buscarem informações de maneira fácil. Segundo o site Ecommerce Brasil (2020), as redes socais crescerá mais de 20% no Brasil até final de 2023.

Assim, com o crescimento excessivo de usuários das redes sociais, o marketing resolveu aderir à essa nova “tecnologia”, já que hoje a maioria das pessoas publicam muito de suas vidas particulares dentro das redes sociais, seus gostos, medos, opiniões políticas, fotos, vídeos, músicas e tudo o que for possível.

De acordo com Lavado (2019), em 2019 a TIC Domicílios afirma que 126,9 milhões de pessoas no Brasil usaram a rede regularmente em 2018. Metade da população rural e das classes D e E agora têm acesso à internet.

Segundo a matéria de Lavado, Winston Oyadomari, o coordenador de pesquisas no Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic), diz que o Brasil tem um crescimento importante e que os países desenvolvidos na América do Norte e Europa já contam com mais de 80% de cobertura de *internet*, deixando o Brasil numa posição intermediária.

Com uma cobertura de mais de 80% de *internet* na América do Norte algumas empresas enxergaram oportunidade de utilizarem o Big Data para coletarem informações disponibilizadas pelos próprios usuários da *internet* para conseguir criar análise de dados, o exemplo disso foi a empresa *Cambridge Analytica* que foi capaz de criar propagandas específicas influenciando o voto nas eleições, caso que ocorreu nas eleições presidenciais norte americana em 2016.

2.1 Quem é Cambridge Analytica

De acordo com Fornasier e Cesar Beck (2020), a *Cambridge Analytica* foi criada em 2013 como parte da SCL (Strategic Communication Laboratories Group) e atuava como uma empresa na qual cujo o seu principal serviço era a análise de dados para fins comerciais ou políticos. A *Cambridge Analytica* tinha como base a junção de *Data Mining, Data Analytics* e Big Data onde o melhor meio de coleta desses dados era através de redes sociais, especificamente o *Facebook.* De acordo com o

Mark Zuckerberg *CEO* (Diretor Executivo) do *Facebook* em 2017, disse que o *Facebook* atingiu 2 Bilhões de pessoas.

> “Pela primeira vez na história do *Facebook,* a rede social reuniu um bilhão de usuários conectados em um mesmo dia, na última segunda-feira (24). Segundo o fundador da empresa, Mark Zuckerberg, o número foi atingido quando “uma em cada sete pessoas se conectaram ao *Facebook* no mesmo período de 24 horas.” (FORBES, 2015).

De acordo com Amanda Imme (2020), o *Facebook* é a segunda rede social mais utilizada do mundo. Muitas pessoas ou empresas utilizam o *Facebook,* entre outras redes sociais, para obter opiniões das pessoas, seus gostos culinários, lugares que mais gostam de passear; podendo assim coletar facilmente qualquer dado no perfil pessoal.

De acordo com o documentário PRIVACIDADE HACKEADA (2020), onde diz que em muitas pessoas acabam compartilhando coisas como, localização, *check-in*, fotos, gosto, time e com isso, em poucos minutos verificando o perfil de algumas pessoas dentro das redes sociais, é possível descobrir e saber o que pode influenciar a pessoa a mudar de ideia em relação às eleições em algum ponto específico. Sendo assim, a *Cambridge Analytica* utilizou a ferramenta para manipular as eleições presidenciais em 2016, onde o gabinete de Donald Trump a contratou para trabalhar com a mineração de dados favorecendo assim as votações para si.

O uso de mineração de dados para fins políticos não é considerado uma prática ilegal, desde que não violem as leis de proteção e privacidade de dados dos usuários.

2.2 Big Data

Conforme Walter Junior (2011), o “Big Data” é um conjunto de dados (*dataset*) no qual o tamanho está além da habilidade de ferramentas típicas de capturar, gerenciar e analisar.

Ainda segundo o autor: "a definição é intencionalmente subjetiva e incorpora uma definição móvel de como um grande conjunto de dados necessita a fim de ser considerado Big Data." (JUNIOR, 2011, v. 14, p. 45)

Ou seja, os conjuntos de dados são tão volumosos que o software de processamento de dados tradicional simplesmente não consegue gerenciá-los. Com

isso, permite que os profissionais de TI trabalhem com informações não-estruturadas a uma grande velocidade.

Para Junior, o Big Data não se define em ser maior do que determinado número de Terabytes, e com a incrementação de dados variados, o Big Data só tende a crescer.

As soluções de Big Data, conforme o CanalTech (2020), foram criadas para lidar com um grande volume e variedade de dados. Isso significa que eles não têm relação entre si e nem uma estrutura definida, por exemplo, *posts* no *Facebook*, *Instagram*, vídeos, fotos, *tweets*, geolocalização, comportamento, ganhos. Além disso, esse tipo de armazenamento de dados com diferentes formatos podem ser usados para satisfazer os clientes de outras maneiras, dando mais conforto a ele em situações incômodas como falta de métricas, gerenciamento e volume.

> "Big Data não trata apenas da dimensão volume, como parece à primeira vista, mas existe também uma variedade imensa de dados, não estruturados, dentro e fora das empresas (coletados das mídias sociais, por exemplo), que precisam ser validados (terem veracidade para serem usados) e tratados em velocidade adequada para terem valor para o negócio. A fórmula é então, Big Data = volume + variedade + velocidade + veracidade, gerando valor." (TAURION, 2013).

Segundo Déborah Oliveira (2015), diariamente são gerados 4,5 quintilhões de *bytes* de dados vindo de variadas fontes de informações. Esses dados vêm de toda parte: troca de mensagens em redes sociais, vídeos, *posts*, registro de pagamento, registro de compra, sinais *GPS* entre outros dados que podem ser armazenados.

2.3 Os 5 Vs do Big Data

Para Fernando Amaral (2018), apesar do Big Data estar ligado à grande quantidade de dados, seu fenômeno não se trata apenas de volume, mas de diversidade, pois são dados gerados de todas as formas e de todas as maneiras. A sua definição é dada por um conjunto de três a cinco "Vs". Os três primeiros se caracterizam em: volume, velocidade e variedade; já os outros dois a mais seguem: veracidade e valor.

Em seu livro "Big Data*:* Técnicas e tecnologias para a extração de valor dos dados", Rosangela Marquesone descreve os "Vs" apresentados no parágrafo acima.

2.3.1 Volume

O volume, segundo Marquesone, é a característica mais significativa no conceito do Big Data, e faz uma referência à dimensão do volume de dados. Em um exemplo, a autora cita uma estatística de 2016 referente ao *Facebook*, que contabilizou “uma média de 1.13 bilhões de usuários, 2.5 bilhões de compartilhamentos e 2.7 bilhões de curtidas diariamente”. Esse exemplo mostra a quantidade de dados necessários para ser arquivado e, desta forma, o que de fato define o volume é a “limitação das ferramentas tradicionais para lidar com determinado volume de dados” (MARQUESONE, 2018).

2.3.2 Velocidade

Ainda citando Marquesone (2018), além do volume e variedade de dados, o Big Data também tem outra propriedade: “a velocidade com que os dados são coletados, analisados e utilizados”. O benefício da velocidade, como exemplo da autora, no varejo é a rapidez para atualizar os valores de acordo com a demanda em tempo real. Assim como afirma a mesma, as empresas que não tem este fator, estão encontrando dificuldades de se manterem competitivas no mercado.

E, além da velocidade no campo da análise, a velocidade do Big Data também se relaciona com a rapidez de que os dados serão gerados. Para ajudar no processamento mais rápido das informações, usando algumas ferramentas como *Apache Spark* e *Databricks* descobriram que transformar os dados de diferentes formatos em um único formato chamado *Parquet* poderia aumentar em até 10 vezes a velocidade que os dados poderiam ser processados.

2.3.3 Variedade

Segundo Marquesone, antes de armazenar informações, é necessário definir sua estrutura, sequência, tamanho e tipos de dados. Os dados são divididos em duas classes: os não-estruturados, como vídeos, imagens, e alguns formatos de texto. Conforme a autora, eles não tem um formato que pode ser facilmente armazenado por tabelas, se tornando complexos para serem processados em “ferramentas tradicionais de armazenamento e gerenciamento de dados”. Já a segunda classe, é conhecida como dados estruturados, que são dados com esquemas rígidos e

adequados para o formato de tabelas.

Com o Big Data*,* mensagens, fotos, vídeos e sons, que são dados não-estruturados, podem ser administrados juntamente com dados tradicionais: “Ou seja, há uma diversidade de dados sendo utilizada por uma variedade de soluções, cada qual com necessidades específicas.” (MARQUESONE, 2018).

Ainda utilizando o argumento da autora, não se sabe explicitamente a quantidade que essa variedade de dados oferece, o que cede erro em muitas fontes de dados e possibilidades de análise.

2.3.4 Veracidade

A veracidade, para a autora, está atribuída ao quão verdadeiro um dado pode ser, sendo correlacionado à confiabilidade dos dados. Por o Big Data estar inserido no contexto com grande quantidade e variabilidade, a inconsistência de dados não é incomum.

Sendo assim, não se pode controlar 100% a gravação desses dados no repositório do Big Data mas, felizmente, existem ótimas ferramentas para refinamento das informações.

2.3.5 Valor

O último V é o que torna Big Data relevante, pois, conforme Marquesone, o valor vai analisar o quão valioso e significativo um dado é referente a uma solução, porque ele vai gerar valor nas informações. Com essas informações sendo armazenadas, tratadas e refinadas é possível gerar valor com os dados, o que permite gerar previsões e pesquisas.

2.4 Documentário Privacidade Hackeada (Netflix 2019)

Para entender um pouco melhor sobre a *CA (Cambridge Analytica)* e todo seu escândalo, será referenciado o filme privacidade hackeada (2019). A história criada pelos diretores Karim Amer e Jehane Noujaim conta sobre os acontecimentos que levaram o *Facebook* ser processado devido a *CA* ter utilizado de forma inapropriada dados pessoais dos usuários do *Facebook,* no documentário é abordado diversos

pontos de vista em relação a toda trama que envolve a *CA* e o *Facebook,* como por exemplo um dos principais nomes que aparece no documentário é do professor David Carroll que foi uma das primeiras pessoas que desconfiou do poder e do que a *CA* fazia com os dados do *Facebook*.

Teve participações importantes também a americana Brittany Kaiser, ex-diretora de desenvolvimento de negócios da *CA* e uma das principais relatoras de como funcionava todo o esquema de coleta de dados da *CA.*

Ainda, Christopher Wylie, que foi um dos principais nomes que fez com que essa enorme coleta e processamento dos dados feito pela *CA* fosse possível mostra o principal objetivo do documentário que é demonstrar como as informações, postagens, fotos e etc, que é colocado nas redes sociais podem ser usados por grandes empresas para obter vantagens econômicas e até sociáveis dentro da sociedade. Houve um exemplo claro de como informações que é julgada "simples" foi utilizada para fazer com que fosse possível a eleição do presidente norte americano Donald Trump.

3. CAMBRIDGE ANALYTICA

A *CA (Cambridge Analytica)* foi criada em 2013 como parte da *SCL (Strategic Communication Laboratories Group)* e atua como serviço de análise de dados para fins comerciais ou políticos. A sede fica em Londres, mas a empresa possui escritórios nos Estados Unidos (Nova York e Washington), Malásia e Brasil. Entre os clientes da *CA* estão Donald Trump e alguns grupos do *Brexit*. O mestre em direito Mateus Fornasier faz um questionamento sobre qual o legado que a *CA* deixou para o mundo.

> "Qual o legado da ocultação, da parte da *CA*, da coleta e do tratamento de dados pessoais e sensíveis que realizou em relação a milhões de eleitores no âmbito de importantes eventos democráticos – tais como o *Referendum* do *Brexit* de 2015 e as Eleições Presidenciais Americanas de 2016?" (FORNASIER, 2020)

A provedora de filmes *Netflix* fez um documentário chamado privacidade hackeada onde explica de vários pontos de vista como foi o funcionamento da *CA* levando em conta visões diferentes de como tudo foi planejado e arquitetado pela *CA*.

Uma das principais peças chaves nesse escândalo todo que envolve a *CA* foi o professor David Carroll, no qual foi uma das primeiras pessoas a desconfiar que seus dados do *Facebook* poderiam estar sendo utilizados de maneira indevida por terceiros. Cátia Rocha (2019), colunista do site português dinheiro vivo, citou David Carroll, Carole Cadwalladr, Brittany Kaiser e Julian Wheatland sendo os principais nomes do documentário.

A *CA* utilizou mapeamento de comportamento para analisar e ver quais eleitores podiam ou não mudar de ideia em relação às eleições. Eles utilizavam questionários no *Facebook* para verificar e descobrir o perfil de cada eleitor nos Estados Unidos.

Com base nisso, mapearam todo o território nacional e verificaram os locais onde havia mais dos "persuasivos" e trabalharam em cima desses para criar toda uma propaganda e unificando a opinião daquela região, de acordo com um artigo escrito por Sofia Caseiro (2020), mestre em Direito Internacional Público e Europeu na Faculdade de Direito da Universidade de Coimbra.

> "A empresa aliava a análise de grandes volumes de dados através de inteligência artificial, com técnicas sociológicas avançadas de análise que permitiam perceber quais eram os eleitores mais influenciáveis e com tendência a mudar de opinião." (CASEIRO, pág. 135-142, 2020)

A *CA* é objeto de investigações criminais tanto nos Estados Unidos como no Reino Unido, de acordo com o site *looptt*, o ministro de segurança nacional *Stuart Young* abriu um processo contra Christopher Wylie.

> "O Ministro da Segurança Nacional, Stuart Young, disse que solicitou oficialmente que uma investigação criminal seja iniciada pelas alegações feitas pelo ex-funcionário da *Cambridge Analytica*, Christopher Wylie, de que o Congresso Nacional Unido (UNC) estava envolvido na violação da privacidade de cidadãos de Trinidad."[1] (LOOPTT, 2019)

Conforme Cadwalladr e Graham-Harrison (2018), a coleta de dados começou quando Aleksander Kogan, um acadêmico da Universidade de Cambrige, desenvolveu um aplicativo chamado "thisisyourdigitallife", que funcionava como um quiz de Facebook. Com isso, Kogan conseguiu que milhares de usuários participassem do teste e, assim, ele tinha acesso a todas as informações que precisava.

> "Os dados foram coletados por meio de um aplicativo denominado essa é sua vida digital (tradução livre para *thisisyourdigitallife*), desenvolvido pelo acadêmico Aleksandr Kogan, separadamente de seu trabalho na Universidade de Cambridge. Por meio de sua empresa, a *Global Science Research* (GSR), em colaboração com a *Cambridge Analytica*, centenas de milhares de usuários foram pagos para fazer um teste de personalidade e concordaram em ter seus dados coletados para uso acadêmico."[2] (CADWALLADR, The Guardian, pág. 1, 2018, tradução nossa)

Até esse ponto, as pessoas tinham aceitado usar o aplicativo, então não era passível de crime ter acesso aos gostos, *likes*, localização e etc. A parte questionável, vista pelas autoras em seu artigo para a *Observer*, é o fato de Kogan repassar as informações coletadas para a *CA*, já que o *Facebook* tem uma política que proíbe o compartilhamento de dados com terceiros, Cadwalladr do *The Guardian* informou que o *Facebook* não aceita as informações dos aplicativos para terceiros e o *Facebook* nega que o acontecimento foi uma falha.

Cadwalladr e Graham-Harrison afirmam que o aplicativo de Kogan tinha uma segunda parte a qual, ao concordarem em usar o aplicativo, o mesmo poderia "entrar"

---

[1] "Minister of National Security Stuart Young said he has officially requested that a criminal investigation be launched into allegations made by former Cambridge we Analytica employee Christopher Wylie that the United National Congress (UNC) was involved in breaching the privacy of Trinidadian citizens"

[2] "The data was collected through an app called thisisyourdigitallife, built by academic Aleksandr Kogan, separately from his work at Cambridge University. Through his company Global Science Research (GSR), in collaboration with Cambridge Analytica, hundreds of thousands of users were paid to take a personality test and agreed to have their data collected for academic use."

nas redes sociais dos seus amigos e minerar os dados dos mesmos. Estima-se que 50 milhões de usuários foram repassados para a *Cambridge Analytica*.

Quando o *Facebook* descobriu a violação desses dados em 2015, ele suspendeu o aplicativo da sua rede social. Kogan, a *Cambridge Analytica* e Christopher Wylie receberam solicitações formais para eliminar os dados coletados. Recentemente, o *Facebook* descobriu que nem todos os dados foram eliminados, o *site* do *The Guardian* revelou que em 2015 o *Facebook* fez um requerimento pedindo para que todos os dados coletados pela CA fossem destruídos.

> *"*O *Facebook* disse que removeu o aplicativo em 2015 e exigiu de todos com cópias dos dados os destruíssem, embora a notificação para Wylie não tenha chegado até o segundo semestre de 2016.*"*[3] (CADWALLADR, *The Guardian*, pág. 4, 2018, tradução nossa)

Christopher Wylie trabalhou para a *Cambridge Analytica* quando estes dados estavam sendo coletados. Ele ajudou a criar modelos de análise comportamentais na qual podiam ser usados para direcionar as ações de cunho político. Caroline Cadwalladr (2018), foi a primeira que mostrou para o mundo o que a *CA* fez.

> "Christopher Wylie, que trabalhou com um acadêmico da Universidade de Cambridge para obter os dados, disse ao Observer: "Destrinchamos o *Facebook* para colher milhões de perfis, assim, construímos modelos para explorar o que sabíamos sobre eles e atingir seus demônios interiores. Essa foi a base sobre a qual a empresa foi construída."[4] (CADWALLADR, 2018, pág. 1, tradução nossa)

Nada mais há a dizer sobre o tema, posto que as palavras da colunista Cadwalladr são suficientes para expôr o assunto.

3.1 Valor dos dados na Internet

O *Internet Archive* é um *site* que tem como principal objetivo preservar alguns dados para gerações futuras, por conta de alguns sites estarem "sumindo" ou apenas ficando defasados. O *Archive* tem como proposta salvar essas páginas para que qualquer pessoa possa acessá-las.

---

[3] "*Facebook* said it removed the app in 2015 and required certification from everyone with copies that the data had been destroyed, although the letter to Wylie did not arrive until the second half of 2016."

[4] "Christopher Wylie, who worked with a Cambridge University academic to obtain the data, told the Observer: "We exploited *Facebook* to harvest millions of people's profiles. And built models to exploit what we knew about them and target their inner demons. That was the basis the entire company was built on".

De acordo com Assunção Duarte (2019), do site Português E-Konomista, o site *Internet Archive* pode ser comparado com a Biblioteca de Alexandria.

> "O *Internet Archive* é uma biblioteca bem diferente da de Alexandria, mas o seu objectivo é muito parecido. Reunir num só sítio tudo o que tem sido criado e publicado *online* desde 1996, ano em que foi fundada pelo americano Brewster Kahle, um engenheiro de computadores e investigador da *internet*. O seu grande objectivo é criar a maior biblioteca digital livre e gratuita do mundo, que armazene tudo o que tem sido publicado *online* nos últimos anos." (DUARTE, Assunção, 2019)

Juntamente com o *Internet Archive*, existe uma ferramenta criada por eles mesmos chamada de *Wayback Machine* (Máquina do tempo), que permite aos usuários acessar algum *site* que já está fora do ar.

Ao digitar a *URL* do *site* é aberto um calendário mostrando as *snapshots* dos *sites*, podendo assim ser possível visualizar o conteúdo de um *site* de 10, 20 anos atrás. De acordo com o próprio *site* do *Wayback Machine* (2020), eles têm cerca de 477 bilhões de páginas da *internet* salvas.

Com base nisso é possível perceber que os dados que circulam na *internet* são de extrema importância, tanto que foi necessário criar um *backup* mundial para a preservação de tais dados. Pereira (2012), colunista do site TecMundo, fez uma publicação mostrando um estudo feito pelo *Internet Archive*, onde foi constatado que o próprio possuía cerca de 10 *petabytes* de dados armazenados.

> "O site *Internet Archive*, responsável por armazenar todo o conteúdo da internet para preservá-lo para gerações futuras, revelou que o arquivo de *backup* da rede mundial de computadores passou da marca de 10 *PetaBytes*. Sabe quanto isso significa em termos mais leigos? O arquivo tem o tamanho de dez milhões de *GB*." (PEREIRA, André, 2012)

Hoje em dia como é esperado, as informações podem ser consideradas como um bem mais valioso, as vezes mais que a pessoa ou a própria empresa. Cada empresa tem alguma informação que para ela é essencial para o funcionamento da mesma.

Não existe um valor real para cada informação, todas são extremamentes importantes e cada uma é utilizada para uma finalidade específica como foi o caso da *Cambridge Analytica* que utilizou de informações postadas pelas pessoas no *Facebook* para criarem uma "inteligência artificial" que processava esses dados e retornava qual o perfil daquela pessoa.

O jornal *The Guardian* citou que, "a comissão eleitoral também está investigando o papel que a *Cambridge Analytica* desempenhou no referendo da União

Europeia (UE)". A comissária de informação Elizabeth Denham, disse

> "Faz parte da nossa investigação em andamento o uso de análise de dados para fins políticos, que foi lançada para avaliar partidos políticos e campanhas. Empresas de análises de dados e plataformas de mídia social no Reino Unido estão usando e analisando as informações pessoais para micro-eleitores-alvo." (The Guardian, 2018)

Até a procuradora geral de Massachusetts chegou comentar no *Twitter* a seguinte fala: "A população merece respostas imediatas do *Facebook* e da *Cambridge Analytica*".

Esse foi o principal motivo pelo qual a União Europeia resolveu mudar suas leis de proteção de dados. Houve um exemplo claro de como informações, o que é postado em redes sociais, fotos e tudo mais tem um valor inestimável no mundo atual que vivemos. Por esse motivo que muitas empresas estão migrando para o mundo digital, pois ele oferece muito mais caminhos e maneiras das empresas lucrarem e crescerem.

3.1.2 Cookies Analíticos

Cookies são micro informações que ficam armazenadas dentro do navegador, que dizem, primeiramente, quais são suas preferências, sem deixar seus dados pessoais. É apenas uma pequena informação que fica guardada para quando você pesquisar algo na internet, já que o próprio navegador vai ler esses "*cookies*" e mostrar facilmente e rapidamente o que você está procurando.

Por exemplo, o site *Education First* (EF), um site de intercâmbio e viagens, utilizam os *Cookies* Analíticos para as seguintes finalidades.

> "Os *cookies* analíticos mostram-nos quais são as páginas mais visitadas no Site, ajudam-nos a registar quaisquer dificuldades que os usuários sintam na navegação no Site, e mostram-nos se a nossa publicidade é eficaz ou não. Isso possibilita ver os padrões globais de uso do Site, em vez da utilização de uma única pessoa." (EF, Education First, 2020)

Dentre os *cookies* existentes, um dos mais utilizados são os cookies analíticos, que tem como principal finalidade ser utilizado anonimamente para para criação e análise de estatísticas, com o principal objetivo melhorar o funcionamento de algum site em específico.

3.2.1 Fake News

Fake News é o termo utilizado atualmente para “notícias falsas” na qual consiste de se pegar alguma informação, mudar o contexto dela e fazer com que ela pareça uma verdade absoluta, como cita Otávio Frias Filho, 2018, que usou como exemplo a viagem à lua.

> “Quando os americanos puseram homens a caminhar sobre a Lua, em 1969, surgiu uma célebre e persistente onda de boatos segundo os quais aquelas imagens haviam sido forjadas em algum estúdio secreto e toda a expedição não passava de um embuste.” (FILHO, Revista USP, v.116, págs. 39-40).

As *fake news* não apareceram nos tempos atuais com o surgimento das tecnologias e os meios de comunicação contemporâneos. De acordo com Filho (2018), durante o século 18 (XVIII) também era muito comumente utilizado durante esse período, levando em conta que as revistas tinham um teor mais anonimato do que atualmente.

> “XVIII, o historiador Robert Darnton mostrou como era infestado de falsidades, plágios, imposturas e calúnias cujos autores ficavam protegidos sob anonimato ou porpseudônimos. De forma semelhante ao que acontece hoje, era comum que os responsáveis por tais abusos escapassem a qualquer punição por estarem fora das fronteiras em que seus escritos haviam sido incriminados.“Ao escrever sobre o fervilhante ambiente panfletário das publicações impressas do século” (FILHO, Revista USP, v.116, págs. 39-40).

Juntamente com as *fake news* existe um outro conceito de manipulação de dados, na qual tem como objetivo pegar uma informação verídica, porém pegar um trecho específico para tirar aquela “fala” para uma coisa fora de contexto.

O colunista Rogério Christofoletti (2018) afirma usando o termo manipulação de dados com uma fala do Erbolato que dissertou o seguinte: “A manipulação da informação acontece quando a notícia tem um tratamento tendencioso ou objetiva mostrar tão somente alguns aspectos.” (ERBOLATO, 1985; apud CHRISTOFOLETTI, 2018).

Pode-se utilizar como exemplo para questão de manipulação de dados o caso que ocorreu nas eleições presidenciais dos Estados Unidos da América em 2016. Quando o embate estava entre Hillary Clinton e o Presidente Donald Trump, um *site* famoso por vazar informações “confidenciais” chamado *WikiLeaks* vazou alguns *email*-s particulares do chefe de campanha da Hillary Clinton. De acordo com Carolina

Canossa (2018), colunista da revista Super interessante, o vazamento ocorreu por volta de março de 2016.

A *fake news* envolvendo a então candidata Hillary Clinton ganhou proporções grandiosas, que até então, pessoas ligadas a Donald Trump teriam contribuído para a propagação da *fake news.* A autora chegou a citar na reportagem a seguinte fala: “Até Michael Flynn, general reformado indicado a conselheiro de segurança nacional do presidente Donald Trump, ajudou a propagar o que aconteceria na Comet Ping Pong em seu Twitter” (Canossa, 2018).

3.5 Engenharia Social

Conforme Marcelo Eiras (2004), Engenharia social é um termo voltado para os estudos das técnicas e práticas utilizadas na obtenção de informações importantes ou sigilosas de uma empresa, por intermédio das pessoas, funcionários e colaboradores de uma corporação ou sociedade. Essas informações podem ser obtidas a partir do emocional das pessoas.

O engenheiro social busca coletar o máximo de informação possível sobre seu alvo, facilitando assim a escolha do método mais efetivo de atingir os usuários.

> “Alguns *hackers* manipulam e enganam. Eles enganam sistemas de computação, fazendo-os pensar que têm autorização que na verdade roubaram; eles praticam engenharia social para manipular as pessoas a fim de atingir seus objetivos.” (MITNICK, pág 31, 2006)

A coleta dessas informações se tornou muito facilitada devido a popularização das redes sociais, onde as pessoas disponibilizam um grande número de informações pessoais, talvez pelo fato de não terem um conhecimento sobre isso, acabam postando e compartilhando algo que contém alguma informação vital em que o atacante pode utilizar para obter vantagem e conseguir aplicar algum golpe relacionado a engenharia social, é tanto que algumas empresas verificam as redes sociais de possíveis pessoas para cargos, conforme é dito pela Doutora e Mestre em Ciências Sociais, Tatiana Martins Almeri.

> “Neste ambiente virtual as agências de emprego aproveitam para observar, monitorar, avaliar e selecionar candidatos a possíveis processos de recrutamento e seleção, considerando o que este expõe em suas redes sociais virtuais. Esta avaliação por parte das agências pode ser positiva ou negativa, tudo depende do conteúdo que é apresentado a ela na página do candidato em potencial.” (ALMERI, pág. 77, 2013)

Após analisar as informações, os engenheiros optaram pelo método que julgam ser mais efetivo em determinados casos.

De acordo com Moretti (2016), do site Administradores, a engenharia social possui uma sequência de passos na qual o ataque ocorre:

- Coleta de informações como números de CPF, datas de nascimento, nome dos pai e familiares, rotinas, *check-in*, que ajudam a estabelecer uma relação entre o atacante e vítima
- Desenvolvimento do relacionamento onde se explora a natureza da confiança das pessoas
- Exploração do ataque o passo onde o engenheiro social utiliza as informações e recursos obtidos

Um dos maiores perigos reside no fato das pessoas exporem muitas informações na *internet*, tanto de maneira direta, postando no *Facebook* ou *Twitter* (fotos de documentos, placas de carro, cartão de crédito), quanto indiretamente através de aplicativos virais que coletam milhares de dados de pessoas no curto período de tempo no qual eles são baixados por milhares de pessoas (*faceapp* e aplicativos que alteram fotos, horóscopos, testes de personalidade).

> “as redes sociais trouxeram consigo a problemática da privacidade e segurança, já que as pessoas colocam informação pessoal e privada e assim perdem a noção de que a “presença virtual” nas redes sociais é real e aporta consequências.” (MACHADO, Claudia, págs. 9-19, 2018)

A engenharia social não é exclusivamente utilizada em informática, ela é uma ferramenta onde se explora falhas humanas em organizações físicas ou jurídicas onde operadores do sistema de segurança da informação possuem poder de decisão parcial ou total ao sistema de segurança da informação seja ele físico ou virtual, Kevin Mitnick deixou um “recado” em seu livro a Arte de Enganar:

> “Quando o invasor de computadores não pode ter acesso físico a um sistema de computador ou à própria rede, ele tenta manipular outra pessoa para fazer isso por ele. Nos casos em que o acesso físico é necessário para o plano, o uso da vítima como representante é melhor ainda do que fazer você mesmo, porque o atacante assume menos risco de ser pego e preso.” (MITNICK, pág. 323, 2003)

Deve-se considerar que as informações pessoais, não documentadas, conhecimentos, saber - não são informações físicas ou virtuais, elas fazem parte de um sistema que possuem características comportamentais e psicológicas na qual a engenharia social passa a ser auxiliada por outras técnicas.

3.6 Manipulação de dados

A engenharia social direcionada é um risco de segurança que se baseia nos próprios usuários. Quanto mais informações e dados os usuários disponibilizam na internet, mais recursos para a realização de ataques os golpistas e engenheiros sociais têm.

Os dados que são expostos vão além de informações pessoais, como endereços e números de telefone e inclui também hábitos e rotinas, interesses, problemas médicos e preferência de serviços. Todos esses dados podem ser manipulados e utilizados contra o usuário em um ataque de engenharia social.

Segundo Townsend (2019), após obter esses dados, seja diretamente através da vitima ou devido a algum vazamento de dados em um sistema vulnerável, o engenheiro social analisa os mesmos e decide qual o melhor método a ser utilizado em cada caso, definindo o que mais vai afetar a vítima aumentando as chances do ataque ter sucesso.

Manipulando os dados obtidos é possível influenciar as escolhas de uma pessoa de várias maneiras, desde mostrar determinadas propagandas até mesmo mostrar informações sobre pessoas de uma maneira que seja conveniente aos responsáveis por essa manipulação, como foi o caso das eleições presidenciais dos Estados Unidos em 2016 envolvendo o ex-presidente Donald Trump.

4. POSSÍVEL CENÁRIO BIG DATA

Conforme a explicação de Krishnan and Eva Tse (2013) relatando sobre as ferramentas e arquitetura na *cloud* utilizada pela *Netflix* para realizar o Big Data, percebe-se que existem muitos cenários e infraestrutura é possível construir usando ferramentas Big Data para chegar onde a empresa *Cambridge Analytica* chegou. Claro que isso é algo muito complexo se comparado com todas as ferramentas disponíveis e o tempo que levaria para construir um imenso mar de dados para realizar o processamento. Mas isso depende de como se pretende que os dados sejam tratados.

Pode ser encontrada ferramentas *Open Source* (Código Aberto) que são softwares gratuitos e muitas ferramentas pagas, aumentando este cenário de criação, como por exemplo, o GCP (Google Plataforma *Cloud*), computação de nuvem da google e AWS (*Amazon Web Service*), plataforma de nuvem da amazon.

Usando a empresa *Cambridge Analytica* como exemplo que usou o processamento de dados a seu favor, foi criado um possível cenário com algumas ferramentas gratuitas onde é feito a leitura de alguns dados em formato de texto. Porém antes de demonstrar na prática será apresentado alguns conceitos e ferramentas.

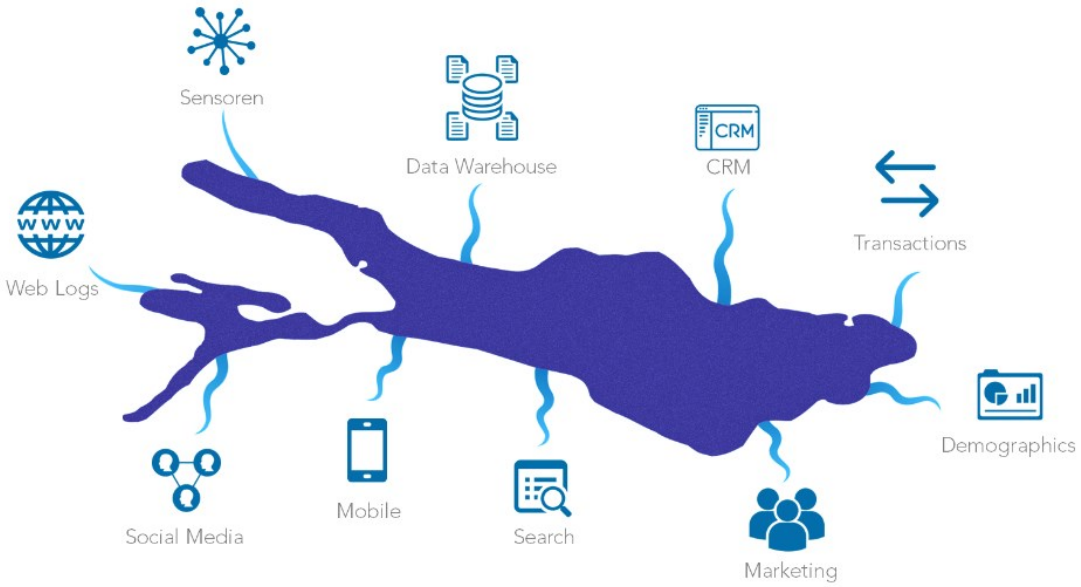
4.1 Conceito Data Lake

O *Data Lake* é conhecido como um conceito no mundo Big Data. Segundo Alrehamy e Walker (2015), *Data Lake* é um enorme repositório onde todos os dados de diferentes formatos, seja ele não-estruturados ou estruturados, são armazenados.

Pense em uma lagoa de informações onde existem vários canais que estão abastecendo essa lagoa, essa é a analogia que pode-se dar para entender melhor o conceito do *Data Lake*.

Na Figura abaixo é demonstrado melhor como um Data Lake funciona e como pode ser construído.

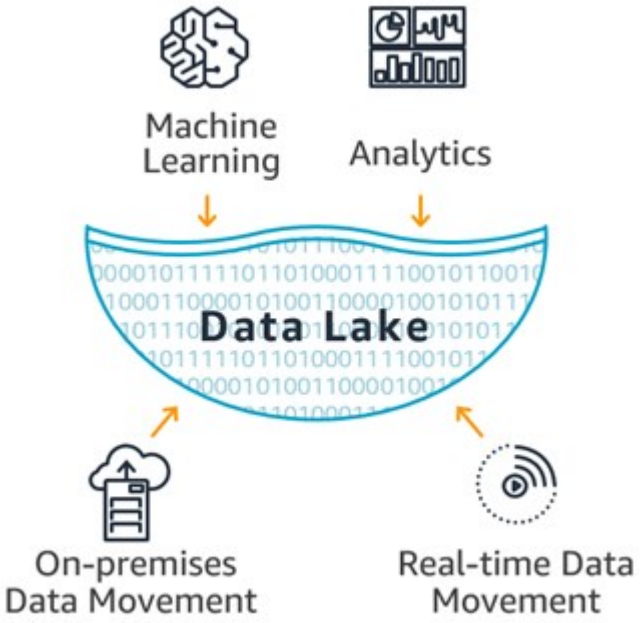
Figura 1 - Demonstração de como é uma *Data Lake.*



Fonte: Lc Systems.

Na Figura 2, é possível ver como o *Data Lake,* com o seu enorme armazenamento de informações, podem ajudar em análises, *machine learning*, pesquisas e previsões.

Figura 2 - Exemplo de como aproveitar os dados do *Data Lake*.



Fonte: AWS Amazon.

4.2 Stream e Batch

O tipo de processamento é muito importante quando existem enormes volumes de dados ou fontes. De acordo com Eran Levy (2019), existem:

- os dados que você processa;
- os dados conforme eles chegam acumula e depois são processados.

O *batch* é usado com frequência para lidar com grandes volumes de dados ou fontes com sistemas legados, que não é possível entregar os dados em fluxos.

Assim como o *batch* o processamento *stream* também pode ser usado para volume de dados maiores, mas o *batch* funciona melhor para análise de dados em tempo real. Porém o *stream* é utilizado para lidar com dados contínuos.

4.3 Ferramenta Open Source.

Segundo Jim Whitehurst, CEO (*Chief Executive Officer*) e presidente d*a Red Hat,* o Código Aberto (*Open Source*) tem ganhado força no mercado de tecnologia nos últimos anos. Atualmente existem milhares de ferramentas e *softwares* de empresas com versões de códigos com a ideia de contribuição. A ideia é que qualquer desenvolvedor ou empresa possa contribuir com a boa funcionalidade do *software*.

Além de *software*, existem muitas ferramentas de processamento de dados 100% *Open Source*, podendo entregar uma infraestrutura totalmente acessível. As ferramentas mais usadas dentro do Big Data são: *Apache Kafka, Hadoop, Apache Spark* e *Apache Hive* ou *Impala*. Essas ferramentas juntas podem formar uma pequena infraestrutura para o início do processamento de dados, armazenamento e consultas.

Na Figura abaixo podemos visualizar uma pequena infraestrutura de processamento de dados 100% *Open Source* feita em *Spark*, *Hive*, *Kafka* e *Hadoop*.

Figura 3 - Exemplo do uso das ferramentas *Open Source* para criação de uma

infraestrutura Big Data.

Fonte: Autoria Própria.

4.4.1 Apache Kafka

O *Apache Kafka* visa proporcionar uma plataforma de alto rendimento e baixa latência para lidar com feeds de dados em tempo real. O *Apache Kafka* também usa o protocolo TCP (*Transmission Control Protocol*) baseado em que é otimizado para eficiência e depende de uma abstração de um agrupamento de mensagens para reduzir a sobrecarga da viagem de ida e volta da rede.

Sua maior missão é oferecer uma infraestrutura de alta disponibilidade, perda zero de mensagens e processamento eficiente de uma só vez. O uso do apache pode ser encontrado na infraestrutura de *e-commerce*, por exemplo, é muito comum a solicitações de pedidos na loja.

4.4.2 Hive

Segundo a documentação da *Microsoft* (2020), o *Apache Hive* é um sistema baseado em *data warehouse* e serve para consultas. *Data Warehouse* é um depósito de dados digitais que tem a ideia de armazenar informações para ajudar a tomar decisões importantes com base nos fatos apresentados.

> “O conceito de *Data Warehouse* surgiu no meio acadêmico em 1980 e significa “armazém de dados”. Trata-se de um sistema que centraliza os dados para análises e BI. Com essa tecnologia, é possível organizar e criar relatórios utilizando históricos.” (TOTVS. 2020)

O *Hive* é usado para leitura de grandes quantidades de dados conjuntos de

dados que residem em armazenamento distribuído usando SQL.

4.4.3 Apache Hadoop

O *Apache Hadoop* é uma ferramenta open source mais usada para realizar armazenamento e processamento de dados usando o *hardware*. O *Hadoop* é um servidor de armazenamento de arquivos com uma estrutura que permite o processamento distribuído de grandes conjuntos de dados em *clusters* de computadores usando modelos de programação simples.

Pode-se confirmar que o *Apache Hadoop* é um sistema de conceito *data lake* está dentro do *Apache Hadoop.*

Logo abaixo será visualizado a página inicial do *Apache Hadoop*, na qual se figura o processo, as métricas e os nós de processamento.

Figura 4 - Página inicial do *Apache Hadoop*.

Fonte: Autoria Própria (2020).

Segundo o *site* Dti Digital, o HDFS (*Hadoop Distributed File System*) servidor de arquivo distribuído, possui o conceito de blocos. Esses blocos normalmente têm tamanho de 64MB. Um arquivo muito grande pode ter blocos armazenados em mais de um servidor.

> *"*O *Hadoop* se tornou o padrão de fato para gerenciar e processar centenas de *terabytes* e *petabytes* de dados. Na *Netflix*, nosso *data warehouse* baseado em *Hadoop* é em escala de *petabyte* e está crescendo rapidamente. No entanto, com a explosão de *Big Data* nos últimos tempos, mesmo isso não é mais uma novidade. Nossa arquitetura, no entanto, é única, pois nos

> permite construir um data warehouse de escala praticamente infinita na nuvem (tanto em termos de dados quanto de poder computacional).” (KRISHNAN, Sriram; TSE, Eva, 2013)

Segundo Krishnan e Tse (2013), afirmam que a arquitetura *Cloud* AWS (*Amazon Web Service*) está usando o *apache* no qual explica com detalhes como utiliza o *Hadoop*, tendo como meio de armazenamento o *Amazon Simple Storage Service*.

4.4.4 Apache Spark

De acordo com Srini Penchikala (2015), o *Apache Spark* é uma estrutura de computação em cluster de código aberto para processamento de dados em tempo real em memória. *Apache Spark* junto com uma linguagem de programação seja ela *Java*, R, *Python* ou *Scala* é capaz de realizar todo o processo de processamento e refinamento dos dados olhando para o conceito de *Data Lake* onde existe um grande repositório de dados. Com o *Spark* é possível se conectar a “lagoa de dados” de dados para realizar todo o refinamento.

De acordo com o Marco Garcia (2020), o principal recurso do *Apache Spark* é o seu mecanismo de alta velocidade de processamento de dados, além disso, é possível transformar os dados em um formato *parquet,* garantindo o processamento dos dados mais rápido. Para demonstrar como o *Spark* trabalha, foi criado uma classe bem simples usando a linguagem de programação *Java* que é possível ler um arquivo de texto, realizar um filtro e salvar o arquivo em *parquet*.

Nesse primeiro passo foi criado uma tabela bem simples de nome para que o *Spark* criado possa ler os dados e realizar o filtro e salvar em um formato *parquet.*

Figura 5 - Exemplo da tabela criada para leitura do *Spark.*

Fonte:: Autoria Própria.

Nesse segundo passo foi realizado a criação da classe “*Spark Hello*” onde é implementado a chamada do construtor do *Spark*. O *Spark Session* é um construtor que tem como objetivo realizar a criação de uma nova sessão do *Spark*.

Figura 6 - Exemplo de chamada para iniciar a sessão do *Spark*.

```java
package murilo.spark.java;
import org.apache.spark.api.java.function.FilterFunction;
import org.apache.spark.sql.Dataset;
import org.apache.spark.sql.SparkSession;

public class SparkHello {
    public static void main(String[] args) {

        SparkSession spark =  SparkSession.builder()
                .master("local")
                .appName("Word Count")
                .getOrCreate();
```

Fonte: Autoria Própria (2020).

No terceiro passo foi introduzindo o caminho do arquivo para que o *Spark* possa realizar a leitura. Logo em seguida, é aplicado um filtro para buscar apenas “Erick” e “Murilo” dentro da coluna *value*. Além disso, foi salvo os dados do filtro em formato *parquet*. E, assim, observa-se todo o código.

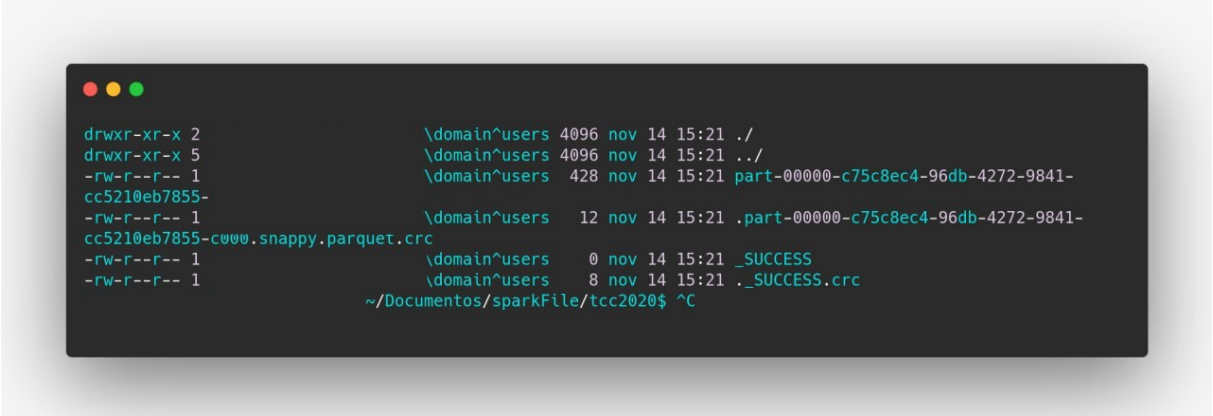
Figura 7 - Exemplo de um filtro utilizando o *Spark Filter* e mostrando a tabela lida.

```
package murilo.spark.java;

import org.apache.spark.sql.Dataset;
import org.apache.spark.sql.SparkSession;

import static org.apache.spark.sql.functions.col;

public class SparkHello {
    public static void main(String[] args) {

        SparkSession spark =  SparkSession.builder()
                .master("local")
                .appName("Word Count")
                .getOrCreate();

        //Realizando a leitura da tabela de nomes
        Dataset<String> nomeTable = spark.read().textFile( "/home/murilo/Documentos/sparkFile" );
        //Mostrando a tabela original
        nomeTable.show();

        //Filtrando João e Murilo dentro da tabela original
        Dataset<String> filtroNameTable = nomeTable.filter(col("value").equalTo("João").
                or(col("value").
                        equalTo("Murilo")));

        //Mostrando a tabela com o filtro adicionado
        filtroNameTable.show();

        //Salvando em formato Parquet
        filtroNameTable.write().format("parquet").save("/home/murilo/Documentos/sparkFile/tcc");


    }
}
```

Fonte: Autoria Própria (2020).

Conforme a Figura 8, verifica-se se o arquivo *parquet* foi salvo no caminho determinado e como os dados ficaram estruturados no formato *parquet*.

Figura 8 - *Parquet* gerado após a execução do *Spark*.



```
drwxr-xr-x 2                    \domain^users 4096 nov 14 15:21 ./
drwxr-xr-x 5                    \domain^users 4096 nov 14 15:21 ../
-rw-r--r-- 1                    \domain^users  428 nov 14 15:21 part-00000-c75c8ec4-96db-4272-9841-
cc5210eb7855-
-rw-r--r-- 1                    \domain^users   12 nov 14 15:21 .part-00000-c75c8ec4-96db-4272-9841-
cc5210eb7855-c000.snappy.parquet.crc
-rw-r--r-- 1                    \domain^users    0 nov 14 15:21 _SUCCESS
-rw-r--r-- 1                    \domain^users    8 nov 14 15:21 ._SUCCESS.crc
                        ~/Documentos/sparkFile/tcc2020$ ^C
```

Fonte: Autoria Própria (2020).

Segundo o site de documentação da *Microsoft*, o *parquet* é um formato de dados assim como o formato CSV (*Comma-separated values*) - valores separados por vírgula -, mas sendo um formato mais eficiente e orientado a colunas do ecossistema *Apache Hadoop*.

A diferença entre os dois formatos é que o *parquet* é um arquivo comprimido, deixando os dados mais leves para serem armazenados em um servidor de arquivos, como o *hadoop*, por exemplo. Além disso, o formato é processado fisicamente em locais de memória.

4.5 Resultado Final

Nos exemplos abaixo como resultado final dos dados filtrados e salvos em formato *parquet* em um diretório.

Na imagem 9, iniciamos o *Apache Spark* em nosso terminal com o comando *Spark-shell*, e lendo o arquivo *parquet* gerado pelo nosso código com o comando:

Figura 9 - Iniciando o *Spark shell* para realizar a leitura do arquivo *parquet*.

Fonte: Autoria Própria (2020).

Na Figura 10, observa-se um comando do *Spark* passando por parâmetro o *parquet* gerado.

Figura 10 - Lendo o arquivo *parquet* gerado na execução do *Spark*.

Fonte: Autoria Própria (2020).

5. CONSIDERAÇÕES FINAIS

Este trabalho apresentou os principais conceitos a respeito do uso do Big Data e como a *Cambridge Analytica* usou a manipulação de dados, engenharia social, *fake news* e processamento de dados para influenciar na eleição dos EUA (Estados Unidos da América) em 2016. Além disso, foi abordado uma pequena introdução de como os dados são importantes nas redes sociais e para as empresas, sendo essencial para evolução da ciência e também para gerenciar métricas e prevenir perdas. Porém, o uso dos dados de forma contrária pode causar grandes prejuízos.

Esse trabalho também abordou o uso das ferramentas *Open Source* (Código Aberto) do Big Data que, de acordo com Jim Whitehurst, CEO (*Chief Executive Officer*) e presidente d*a Red Hat,* diz que o código aberto tem ganhado força no mercado de tecnologia nos últimos anos. O exemplo são as ferramentas*, Apache Hive, Apache Spark, Apache Hadoop* e *Apache Kafka,* todas essas ferramentas podem ser encontradas gratuitamente, permitindo qualquer programador criar seu próprio coletor de dados.

Foi efetuada a análise dos dados usando a linguagem *Java* e o *Apache Spark* que é responsável pela leitura, processamento e tratamento dos dados. No exemplo do capítulo 4 é possível visualizar que o *Spark* foi usado para realizar um filtro em um arquivo de texto com dados contendo diversos nomes, logo em seguida foi salvo em um diretório em formato *parquet.*

O objetivo final do trabalho foi mostrar como os dados são informações importantes podendo impactar grandes decisões políticas. Também foi mostrar como é possível criar pequenos coletores que realizam leituras e interpretações dos dados de forma fácil, usando apenas ferramentas gratuitas.

Sugere-se para trabalhos futuros, estudar melhores maneiras de criar ou aprimorar o coletor *Spark,* melhorando o processamento dos dados e garantido boas práticas de seu uso*,* assim como novas implementações de ferramentas *Open Source* (Código Aberto) para acrescentar na eficiência do processamento de dados e na independência do coletor.

REFERÊNCIAS

ABOUT the Internet Archive. *S.L*., 201-?. Disponível em: https://archive.org/about/. Acesso em: 6 out. 2020.

AMARAL, Fernando. **Introdução à Ciência de Dados:** mineração de dados e big data. 01. ed. *S.L*.: Alta Books, 2018. cap. Introdução, p. 4-18. Disponível em: https://books.google.com.br/books?hl=pt-BR&lr=lang_pt&id=hAlVDQAAQBAJ&oi=fnd&pg=PR13&dq=BIG+DATA&ots=hG5m4u9vA2&sig=bmqnz9-ov8gor-omlt2Tt0-rKBM#v=onepage&q=BIG%20DATA&f=false. Acesso em: 14 nov. 2020.

ALMERI, Tatiana Martins *et al*. O uso das redes sociais virtuais nos processos de recrutamento e seleção. **ECCOM**, São José dos Campos, v. 4, n. 8, p. 77-94, dez. 2013.

ALREHAMY, H.; , WALKER C. (2015). **Lago de dados pessoais com atração de gravidade de dados.** Na 5ª conferência internacional IEEE sobre big data e computação em nuvem (BDCloud 2015), Dalian, China, IEEE computer society washington, vol. 88, pp. 160–167. Disponível em: https://doi.org/10.1109/BDCloud.2015.62 .

AWS Amazon. **O que é um data lake?** Disponível em: https://aws.amazon.com/pt/big-data/datalakes-and-analytics/what-is-a-data-lake/ Acesso 14 nov. 2020.

BATTAGLIA, Rafael. **Como identificar e combater fake news?**. *S.L*, 8 out. 2018. Disponível em: https://super.abril.com.br/sociedade/como-identificar-e-combater-fake-news. Acesso em: 7 out. 2020.

CADWALLADR, Carole; GRAHAM-HARRISON, Emma. **Revealed:** *50 million Facebook profiles harvested for Cambridge Analytica in major data breach***.** The Guardian, mar. 2018. Disponível em: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election. Acesso em: 29 nov. 2020.

CANALTECH. **O que é big data?** Disponível em https://canaltech.com.br/big-data/o-que-e-big-data/ Acesso em: 25 de Agosto de 2020.

CANALTECH. **Big Data: os cinco Vs que todo mundo deveria saber?.** Disponível em https://canalt ech.com.br/big-data/Big-Data-os-cinco-Vs-que-todo-mundo-deveria-saber Acesso em: 25 de Agosto de 2020.

CANOSSA , Carolina. Pizzagate: **O escândalo de fake news que abalou a campanha de Hillary.** S.L, 13 abr. 2018. Disponível em: https://super.abril.com.br/mundo-estranho/pizzagate-o-escandalo-de-fake-news-que-abalou-a-campanha-de-hillary/. Acesso em: 3 nov. 2020.

CASEIRO, Sofia. **O Impacto da Inteligência Artificial na Democracia**. *In*: IV CONGRESSO INTERNACIONAL DE DIREITOS HUMANOS DE COIMBRA: UMA VISÃO TRANSDISCIPLINAR, 2019, Coimbra. **ANAIS DE ARTIGOS COMPLETOS** [...]. Brasil: Edições Brasil; Editora Fibra; Editora Brasílica, 2020. p. 135-142.

CHRISTOFOLETTI, R. **Padrões de manipulação no jornalismo brasileiro:** fake news e a crítica de Perseu Abramo 30 anos depois. RuMoRes, *[S. l.]*, v. 12, n. 23, p. 56-82, 2018. DOI: 10.11606/issn.1982-677X.rum.2018.144229. Disponível em: https://www.revistas.usp.br/Rumores/article/view/144229. Acesso em: 3 nov. 2020.

CNSEG. **Como o big data e a psicometria ajudaram Trump a vencer as eleições americanas?**. 16 de Fevereiro de 2017. Disponível em https://cnseg.org.br/noticias/como-o-big-data-e-a-psicometria-ajudaram-trump-a-vencer-as-eleicoes-americanas.html. Acesso em: 25 de ago. 2020.

COSSETTI, Cruz Melissa. **Facebook chega a 2 bilhões de usuários.** Techtudo, 27 Junho de 2017. Disponível em https://www.techtudo.com.br/noticias/2017/06/facebook-chega-a-2-bilhoes-de-usuarios.ghtml. Acesso em: 20 de ago. 2020.

DUARTE, Assunção. **Internet Archive:** *Uma biblioteca digital gigante e gratuita***.** E-konomista. 16 de Abril de 2019. Disponível em https://www.e-konomista.pt/inter net-archive/ Acesso em: 06 de out. 2020.

ECOMMERCE BRASIL, 2020. **Usuários de redes sociais crescerão em mais de**

20% no Brasil até final de 2023. Ecommerce Brasil 20 de fevereiro de 2020. Disponível em: <https://www.ecommercebrasil.com.br/noticias/usuarios-de-redes-sociais-crescerao-em-mais-de-20-no-brasil-ate-final-de-2023/#:~:text=Para%20a%20popula%C3%A7%C3%A3o%20brasileira%2C%20a,do%20per%C3%ADodo%20de%20tr%C3%AAs%20anos>. Acesso em: 13 nov. 2020.

EDUCATION First. **Política de Cookies.** Disponível em: https://www.ef.com.br/legal/cookie-policy/ Acesso em: 18 nov. 2020.

EIRAS, Marcelo Coradassi. Engenharia Social. **Engenharia Social e Estelionato Eletrônico**. 2004. Monografia (Graduação "Lato Sensu" em Segurança de Informações na Internet) - Universidade Federal do Estado do Rio de Janeiro, RJ, 2004. p. 40. Disponível em: https://docplayer.com.br/983029-Engenharia-social-e-estelionato-eletronico.html. Acesso em: 5 nov. 2020.

FOLHA DE SÃO PAULO. **'Sou bode expiatório', diz criador do app usado para obter dados no Facebook.** 2018. Acesso em: 21 de março de 2020 Disponível em: https://www1.folha.uol.com.br/mercado/2018/03/sou-bode-expiatorio-diz-criador-do-app-usado-para-obter-dados-no-facebook.shtml

FORNASIER, M., & BECK, C. (2020). **CAMBRIDGE ANALYTICA: ESCÂNDALO, LEGADO E POSSÍVEIS FUTUROS PARA A DEMOCRACIA.** Revista Direito Em Debate, 29(53), 182-195. Disponível em: https://doi.org/10.21527/2176-6622.2020.53.182-195

FORBES. **Facebook atinge 1 bilhão de pessoas conectadas em um único dia.** 28 de agosto de 2015. Disponível em: https://www.forbes.com.br/colunas/2015/08/facebook-atinge-1-bilhao-de-pessoas-conectadas-em-um-unico-dia/. Acesso em: 14 nov. 2020.

FILHO, Frias, O. (2018). **O que é falso sobre fake news.** Revista USP, (116), 39-44. Disponível em: https://doi.org/10.11606/issn.2316-9036.v0i116p39-44 Acesso em: 19 nov. 2020.

G1, 2017. **Facebook atingiu 2 milhões de usuários.** G1. 27 de Junho 2017. Disponível em: https://g1.globo.com/tecnologia/noticia/facebook-atinge-os-2-bilhoes-

de-usuarios.ghtml. Acesso em: 08 out. 2020.

GARCIA, Marco. **Spark: Saiba mais sobre esse poderoso framework.** CETAX. agosto 8, 2020. Disponível em: https://www.cetax.com.br/blog/conheca-mais-sobre-o-framework-apache-spark/#:~:text=benef%C3%ADcios%20do%20Spark%3F-,VELOCIDADE,computa%C3%A7%C3%A3o%20mem%C3%B3ria%20e%20outras%20otimiza%C3%A7%C3%B5es. Acesso em: 19 nov. 2020.

IMME, Amanda. **As 10 redes sociais mais usadas no Brasil.** Resultados Digitais 21 de janeiro de 2020. Disponível em: https://resultadosdigitais.com.br/blog/redes-sociais-mais-usadas-no-brasil/#:~:text=2.%20Facebook&text=E%2C%20claro%2C%20segue%20sendo%20a,de%20%C3%8Dndia%20e%20Estados%20Unidos. Acesso em: 19 nov. 2020.

JUNIOR, Walter Teixeira Lima. Jornalismo computacional em função da “Era do Big Data”. Revista Líbero, São Paulo, v. 14, n. 28, p. 45-52, 15 dez. 2011. Disponível em: http://seer.casperlibero.edu.br/index.php/libero/article/view/329/303. Acesso em: 5 nov. 2020.

KRISHNAN, Sriram; TSE, Eva. **Hadoop Platform as a Service in the Cloud.** S.L, 2020. Disponível em: https://netflixtechblog.com/hadoop-platform-as-a-service-in-the-cloud-c23f35f965e7. Acesso em: 6 nov. 2020.

LAVADO, Thiago. **Uso da internet no Brasil cresce, e 70% da população está conectada.** G1. 28 de Agosto de 2019. Disponível em: G1 https://g1.globo.com/economia/tecnologia/noticia/2019/08/28/uso-da-internet-no-brasil-cresce-e-70percent-da-populacao-esta-conectada.ghtml Acesso em: 20 de Agosto de 2020.

LC, Systems. **Data lake em vez de data swamp.** Disponivel em https://www.lcsystems.de/losungen/data-analytics-new/data-lake/ acesso Acesso em: 14 nov. 2020.

LOOP News (2020). **Iniciada investigação criminal sobre as relações UNC-**

Cambridge Analytica. LOOP News. 13 de novembro de 2019. Disponível em: https://www.looptt.com/content/criminal-investigation-launched-unc-cambridge-analytica-tie Acesso em: 19 nov. 2020.

LEVY, Eran. **Processamento em lote, fluxo e microlote: uma folha de referências.** Disponivel em: https://www.upsolver.com/blog/batch-stream-a-che at-sheet Acesso em 15 nov. 2020.

MACHADO, Claudia**. O lado negro das redes sociais.** Disponível em**: https://dialnet.unirioja.es/servlet/articulo?codigo=6750333 .** Acesso em: 20 nov. 2020.

MARQUESONE, Rosangela. **Big Data:** *Técnicas e tecnologias para extração de valor dos dados.* São Paulo: Casa do Código, 2016. Disponível em: https://books.google.com.br/books?hl=pt-BR&lr=lang_pt&id=cbWlDQAAQBAJ&oi=fnd&pg=PT2&dq=big+data+5+v%27s&ots=6mXop7l6J8&sig=qSdH6jHJP_aSHAUoRtkWGodYPCY#v=onepage&q=big%20data%205%20v‘s&f=false. Acesso em: 14 nov. 2020.

MICROSOFT, 2020. **Saiba mais sobre o Apache Hive e o HiveQL no Azure HDInsight?** Disponível em: https://docs.microsoft.com/pt-br/azure/hdinsight/had oop/hdinsight-use-hive#:~:text=Apache%20Hive%20%C3%A9%20um%20sistma,consultas%20e%20an%C3%A1lise%20de%20dados Acesso em: 14 nov. 2020.

MICROSOFT, 2020. **Arquivo Parquet.** Microsoft. 18 de Agosto de 2020. Disponivel em: https://docs.microsoft.com/pt-br/azure/databricks/data/data-sources/read-parquet#:~:text=O%20Apache%20parquet%20%C3%A9%20um,do%20que%20CSV%20ou%20JSON. Acesso em: 18 nov. 2020.

MITNICK, Kevin. **A arte de invadir.** 2. ed. rev. São Paulo: Pearson Education do Brasil, 2005. 245 p. v. 1. ISBN 85-7605-055-2.
MITNICK, Kevin. **A arte de enganar**. 01. ed. rev. São Paulo: Pearson Education do

Brasil, 2003. 588 p. ISBN 85-346-1516-0.

MORETTI, C. Santos. **Atenção com os golpes dos engenheiros sociais. Administradores 15 de Dezembro de 2016** Disponível em: https://administradores.com.br/artigos/atencao-com-os-golpes-dos-engenheiros - sociaishttps://administradores.com.br/artigos/atencao-com-os-golpes-dos-engenheiros-sociais Acesso em 20 nov. 2020.

OLIVEIRA, Déborah. **Analytics:** *comece pequeno e depois amplie, aconselha IBM. S.L*, 2015. Disponível em: https://itforum.com.br/noticias/apesar-de-avancos-nuvem-ainda-gera-debates-nas-empresas/. Acesso em: 13 nov. 2020.

PEREIRA, Luiz André. **Site revela tamanho atual do arquivo com todo o conteúdo da internet.** TecMundo. 29 de Outubro de 2012. Disponível em https://www.tecmundo.com.br/internet/31932-site-revela-tamanho-atual-do-arquivo-com-todo-o-conteudo-da-internet.htm Acesso em 06 de Outubro de 2020.

PENCHIKALA, Srini. **Big Data com Apache Spark - Parte 1: Introdução.** Disponível em: **https://www.infoq.com/br/articles/apache-spark-introduction/** Acesso em 14 nov. 2020.

PRIVACIDADE HACKEADA. Direção: Jehane Noujaim, Karim Amer. Produção: Karim Amer, Jehane Noujaim, Pedro Kos, Judy Korin, Geralyn Dreyfous. Roteiro: Karim Amer, Erin Barnett, Pedro Kos. [*S. l.*]: Netflix, 2019. Disponível em: https://www.netflix.com/br/title/80117542?source=35. Acesso em: 14 mar. 2020.

ROCHA, Cátia. **The Great Hack, o documentário que mostra o poder dos dados que dá às empresas.** Dinheiro vivo. 23 de Julho de 2019. Disponível em: https://www.dinheirovivo.pt/empresas/the-great-hack-o-documentario-que-mostra-o-poder-dos-dados-que-da-as-empresas-12808785.htm Acesso em: 2020 nov. 2020.

TAURION, Cezar. Big data. Rio de Janeiro: Brasport Livros e Multimídia Ltda., 2013.

TECMUNDO: **SITE revela tamanho atual do arquivo com todo o conteúdo da**

internet.
TECMUNDO. 29 Outubro 2012. Disponível em: <https://www.tecmundo.com.br/internet/31932-site-revela-tamanho-atual-do-arquivo-com-todo-o-conteudo-da-internet.htm.> Acesso em: 6 out. 2020.

TOTVS. 2020. **Quais as vantagens de um Data Warehouse?** 14 de Abril 2020. Disponível em: https://www.totvs.com/blog/negocios/data-warehouse/ Acesso em: 14 nov. 2020.

TOWNSEND, Kevin. **Engenharia social: não se trata apenas de golpes de phishing.** Blog Avast. 3 de Maio de 2019. Disponível em: https://blog.avast.com/pt-br/social-engineering-hacks Acesso em: 13 de Setembro de 2020.

WHITEHURST, Jim. **O Estado do Open Source Corporativo.** Red Hat. Disponível em: https://www.redhat.com/cms/managed-files/rh-enterprise-open-source-report-detail-f21756-202002-a4-ptbr.pdf Acesso em: 14 nov. 2020

WAYBACK Machine. *S.L.*, 201-?. Disponível em: http://web.archive.org/. Acesso em: 6 out. 2020.