Centro Universitário Carioca

Ciência da Computação

NELSON RODRIGUES DOS SANTOS JUNIOR

Análise de tecnologias em Big Data.

Rio de Janeiro

2022

Nelson Rodrigues dos Santos Junior

Análise de tecnologias em Big Data.

Trabalho de conclusão de curso apresentado ao curso de Graduação em Ciência da Computação da UNICARIOCA, para o título de Bacharel em Ciência da Computação.

Orientador: Marcelo Perantoni

Rio de Janeiro – RJ

2022

Nelson Rodrigues dos Santos Junior

Análise de tecnologias em Big Data.

Trabalho de conclusão de curso apresentado ao curso de Graduação em Ciência da Computação da UNICARIOCA, para o título de Bacharel em Ciência da Computação.

Aprovado em: Rio de Janeiro/RJ, _____/_____/_____ pela seguinte banca examinadora.

_____

Professor Marcelo Perantoni. D. Sc. – Orientador
Centro Universitário Carioca

_____

Professor Fabio Henrique Silva. M. Sc. – Professor Convidado
Centro Universitário Carioca

_____

Professor André Luiz Avelino Sobral. M. Sc. – Coordenador de Curso
Centro Universitário Carioca

Resumo

O gerenciamento e a manipulação da informação têm seu crescimento atrelado a facilidade de acesso e inserção a informação que possuímos, onde uma quantidade massiva de dados gerados pela internet é gerada e armazenada. Nesse contexto, foi necessária uma evolução na manipulação de dados, a era de Big Data, que passou a oferecer soluções a problemas outrora irresolúveis devido às limitações das ferramentas disponíveis, soluções as quais são compostas de tecnologias inovadoras para o tratamento, armazenamento e manipulação dos mais variados formatos de dados de forma eficiente e acessível. Isto posto, no presente trabalho são apresentadas as tecnologias mais utilizadas para o armazenamento, tratamento e visualização dos dados, bem como outras para integração de tecnologias e manipulação de contêineres em um contexto de big data.

Palavras-chave: Big Data; Análise; Tecnologia;

Abstract

The management and manipulation of the information have their growth linked to the facility of access and insertion of the information that we possess, where a massive amount of data created from the internet is generated and stored. In this context, an evolution in data manipulation was necessary, the age of Big Data, which began to offer solutions to problems once irresolvable because of limitations of the available tools, solutions which are composed of breakthrough technologies for the processing, storing and manipulation of the most various data types in an effective and affordable way. Therefore, this work presents the most commonly used technologies for the storing, processing and viewing of data, as well as others for integration of technologies and container manipulation in a big data context.

Keywords: Big Data; Analysis; Technology;

Lista de Figuras

Sumário

1. INTRODUÇÃO

O gerenciamento e manipulação da informação tem seu crescimento atrelado a facilidade de acesso e inserção a informação que possuímos, em que uma quantidade massiva de dados gerados pela internet seja em vídeo, texto, imagem, ou dentre outros formatos, são processados e armazenados para uso ou visualização, e é através de ferramentas de big data que esse tratamento é feito.

1.1 Contexto

As duas grandes guerras mundiais geraram grandes avanços tecnológicos, mas foi após o fim da segunda guerra, que o primeiro computador eletrônico e digital, o ENIAC, foi desenvolvido, uma ferramenta para processamento de informação que evoluiu também para um meio de comunicação, em 1969 com a criação da ARPANET, uma intranet entre centros de pesquisa. Conforme o meio de comunicação digital evoluía, o poder de processamento foi ampliado e mais informação pôde ser processada e trafegada na rede.

No cenário atual, no qual há uma quantidade massiva de informações sendo trafegadas diariamente em diversos formatos, sendo necessário o seu processamento e armazenamento em tempo hábil, e para isso são utilizadas ferramentas específicas para o tratamento desse grande volume.

1.2 Objetivos

Esse trabalho possui o objetivo de desenvolver um estudo sobre big data, seus fundamentos, suas ferramentas, e onde os dados são gerados, bem como desenvolver uma pesquisa acerca de suas tecnologias.

Utilizando artigos, textos, experiências, citações e dados como base, tem o intuito de informar ao leitor desde uma base média de conhecimento do assunto, até aspectos de ferramentas, que são mais usuais para interessados no campo da pesquisa.

1.3 Justificativa

Conforme o mundo digital vem se expandindo, cada vez mais pontos de informação são criados, e o uso dessas informações passam a ser também de dispositivos digitais, com a expansão da internet das coisas. O tratamento da informação é uma necessidade, e quantidade não pode ser um impeditivo, com isso, qualquer entidade que utilize de serviços relacionados, eventualmente terão de recorrer a soluções em *big data*.

Nesse contexto o trabalho mostrará como o estudo do tema pode ser aplicado em qualquer área, bem como os aspectos positivos acerca de *big data* que tornaram suas soluções amplamente difundidas.

1.4 Estrutura do Trabalho

O trabalho é constituído de uma definição do tema, seus pilares e contexto histórico. No capítulo 2 é desenvolvido a pesquisa sobre big data, bem como suas características, benefícios, exemplos de uso, seguindo dos processos que compõem o assunto e por fim a pesquisa acerca das tecnologias utilizadas em soluções de big data, suas características, benefícios, seus casos de uso e por quem estão sendo utilizadas, finalizando com uma lista de outras tecnologias que também estão inseridas no contexto de big data.

2. REFERENCIAL TEÓRICO

2.1 Big Data

O termo *Big Data* se refere a coleta, armazenamento, tratamento e manipulação de uma massiva quantidade de dados, que não seriam possíveis através de soluções convencionais, utilizando de ferramentas específicas que processam um grande volume de dados e os disponibilizam de forma organizada para uso, que podem ser utilizadas para resolver problemas de negócio que antes não poderiam ser resolvidas devido a limitação de ferramenta (Morais et al., 2018).

Figura 1 - Big Data.
Fonte: https://medium.com/techiepedia/introduction-to-big-data-a7aa24bf17c5.

Um dos modelos mais utilizados para processamento de big data é o *MapReduce*, um modelo de programação definido por um conjunto de bibliotecas para o processamento de um grande volume de dados, tendo como características as operações de mapear (análise e classificação) e reduzir (agrupar os dados já filtrados), operando de modo paralelo e recursivo em diversos volumes de dados, com o objetivo de chegar a um resultado específico que será utilizado para a tomada de decisão (EVEO, 2022).

Com o advento da internet, assistir vídeos, escutar música, mandar mensagens, fazer uma pesquisa são tarefas que se tornaram cada vez mais simples. Com isso, a quantidade de informação trafegada aumentou, além de ser processada e apresentada, como uma pesquisa no google por exemplo, e essa grande variedade, em grande volume e grande velocidade

trafegando e tendo que ser processadas, caracterizam os três Vs de big data como demonstrado na figura 2.

Volume	A quantidade de dados importa. Com o big data, você terá que processar grandes volumes de dados não estruturados de baixa densidade. Podem ser dados de valor desconhecido, como feeds de dados do Twitter, fluxos de cliques em uma página web ou em um aplicativo para dispositivos móveis, ou ainda um equipamento habilitado para sensores. Para algumas empresas, isso pode utilizar dezenas de terabytes de dados. Para outras, podem ser centenas de petabytes.
Velocidade	Velocidade é a taxa mais rápida na qual os dados são recebidos e talvez administrados. Normalmente, a velocidade mais alta dos dados é transmitida diretamente para a memória, em vez de ser gravada no disco. Alguns produtos inteligentes habilitados para internet operam em tempo real ou quase em tempo real e exigem avaliação e ação em tempo real.
Variedade	Variedade refere-se aos vários tipos de dados disponíveis. Tipos de dados tradicionais foram estruturados e se adequam perfeitamente a um banco de dados relacional. Com o aumento de big data, os dados vêm em novos tipos de dados não estruturados. Tipos de dados não estruturados e semiestruturados, como texto, áudio e vídeo, exigem um pré-processamento adicional para obter significado e dar suporte a metadados.

Figura 2 - três Vs Big Data.
Fonte: Oracle, 2022.

Nos últimos anos, surgiram mais dois Vs (Morais et al., 2018):

- Veracidade: é necessário obter dados verídicos de acordo com a realidade, que é bem alinhado ao conceito de velocidade pela necessidade constante de análise em tempo real.
- Valor: importância que os dados possuem, de modo que o processo de *Big Data* irá solucionar questionamentos de um negócio.

O termo big data vem ganhando bastante popularidade na última década, porém suas origens remontam as décadas de 60 e 70, com os primeiros data centers e o desenvolvimento de bancos de dados relacionais (ORACLE, 2022).

Em 2005, foi desenvolvida a Hadoop, uma das tecnologias de processamento para big data, auxiliou em seu crescimento devido à redução do custo para armazenamento e facilidade de manipulação dos dados. Em 2009, o Spark, também uma tecnologia de processamento de dados, mas que efetua o processamento na memória contrastando do Hadoop que é em disco, começou a ser desenvolvido e em 2010 virou um framework de código aberto, com isso teria

duas opções de processamento, sendo uma com alto armazenamento e outra com alta velocidade. A partir de 2005, o volume de dados aumentou massivamente, com o crescimento da Internet das Coisas (IoT), com mais dispositivos e objetos conectados à internet; e o surgimento de machine learning, não só humanos, mas usuários máquinas também contribuíram para produção de dados (ORACLE, 2022).

O potencial para soluções em Big Data é gigante e bem abrangente, podendo beneficiar ramos de negócios completamente diferentes. Através do armazenamento de dados em diversos formatos e em grandes quantidades, seu processamento gera uma base informacional confiável e eficiente que permite uma predição e descoberta de fatores utilizando inteligência computacional, gerando conteúdos relevantes ao negócio (Aquarela, 2022).

Casos de Uso em Big Data:

- Netflix - a netflix implementa, através de modelos de data analytics, um refinamento dos padrões de consumo de seus usuários, usando essas informações para personalizar suas recomendações que, de acordo com a empresa, em torno de 75% das atividades dos usuários giram em torno de conteúdos personalizados recomendados pela plataforma.
- Google - utiliza big data para otimizar e refinar suas buscas e seus algoritmos de propaganda, utiliza em suas pesquisas para obter resultados potencialmente úteis e algoritmos de machine learning para garantir confiabilidade dos dados, além de continuar desenvolvendo aplicações e serviços que possuem algoritmos de big data.
- Wal-Mart - utiliza big data para analisar as grandes informações de suas operações, provendo dados de suas lojas, farmácias e centros de distribuição, de modo a aumentar a eficiência, personalizar a experiência do usuário, escolher a melhor rota de suprimento, entre outros.
- Uber - usa dados pessoais do usuário para monitorar os serviços mais utilizados, analisar padrões e determinar mudanças nos preços e no oferecimento de promoções.
- Centros de controle e prevenção de doenças - através de um histórico de dados, o Google compara as pesquisas com base em áreas geográficas que possuem um histórico de doenças, como por exemplo a gripe, e através desses dados, é possível otimizar o trabalho dos centros de controle e prevenção de doenças.

2.1.1 Tecnologias de Big Data

Big Data classifica não só um grande volume de dados, mas também inclui as ferramentas de análise (*analytics*) e as tecnologias de infraestrutura, que fazem o tratamento e armazenamento dos dados. As tecnologias de Big Data possibilitam otimizar e prever os dados através da criação de modelos estatísticos (Morais et al., 2018).

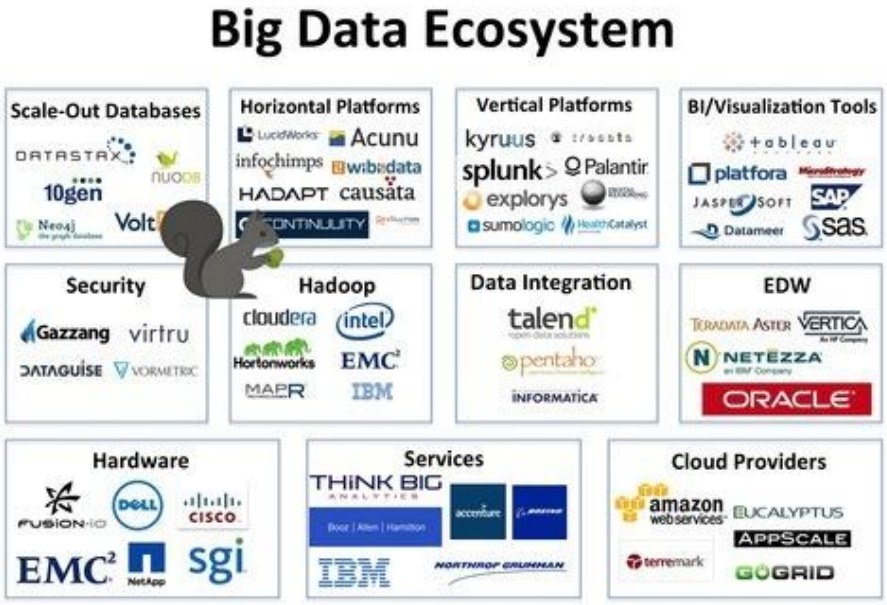

Figura 3 - Big Data ecosystem.
Fonte: bigdatanews.com/.

De acordo com Morais et al., 2018:

- Hadoop – Software de computação distribuída em linguagem java voltada para a manipulação de grandes volumes de dados e com suporte a tolerância de falhas.
- Map Reduce – Framework da google com suporte a computação paralela em grandes volumes de dados em *Clusters* de computadores.
- Linguagens script – linguagens de programação adequadas como python, e R.
- Visual Analytics – método de análise de dados com saída em formato visual.
- Processamento de linguagem natural (PLN) – inteligência artificial para geração de textos.
- In-memory analytics – processamento de *Big Data* realizado na memória, visando velocidade de análise, evitando que os dados sejam armazenados em disco.

De acordo com Lorenzi, 2022:

1. Análise Preditiva: uma tecnologia que ajuda a descobrir, avaliar, otimizar e implantar modelos preditivos através de uma IA e machine learning em bases de big data, auxiliando no desempenho, redução de riscos e vantagem competitiva ao aprender com o passado, visualizar o presente e prever o futuro.
2. Banco de Dados NoSQL: diferente dos bancos de dados relacionais, eles estão em crescimento exponencial, oferecendo um esquema dinâmico, possibilitando uma personalização além de mais flexibilidade e escalabilidade, muito visado ao se armazenar dados de big data.
3. Ecossistema Hadoop: o framework foi desenvolvido para armazenar e processar de modo distribuído com alta velocidade e baixo custo, através da utilização de um modelo de programação simples em um ambiente distribuído. Sempre sendo adotado como tecnologia de big data.
4. Stream Analytics: é uma análise de dados em tempo real de grandes "pools de dados", de inserção em tempo real, sendo utilizado de consultas contínuas ou "chamadas de fluxo de eventos". Auxilia na descoberta de padrões ocultos, correlações e outras respostas quase que em tempo real, indicado para diversas ações ágeis.
5. Docker: é uma solução que implementa o conceito de conteinerização, que visa desenvolver aplicações que serão desenvolvidas em um ambiente isolado, que pode ser replicado em outras máquinas em diferentes plataformas, com todos os recursos necessários, facilitando a sua implementação, seu escalonamento e sua execução, sem que outras variáveis ou recursos de diferentes sistemas operacionais possam interferir. Ideal para todos os aplicativos necessários com a alocação mínima de recursos.
6. Kubernetes: solução da Google de código aberto, com ideia semelhante a Docker, que possibilita o gerenciamento de múltiplos contêineres, oferecendo liberdade para o desenvolvimento em seu próprio "cluster local".
7. Data Lake: um "lago de dados", um repositório que armazena qualquer formato de dados, sendo estruturado, semiestruturado ou não estruturado, como por exemplo, banco de dados relacionais, fotos e dados de sensores, respectivamente. Possibilita o armazenamento dos dados antes de serem processados, possibilitando análise e manipulação, para visualização e até para uso em aplicações ágeis. Seu grande diferencial se deve a realizar diversas análises por meio de arquivos de log, dados

de mídia social e streaming, possibilitando uma tomada de decisão quase que em tempo real.

8. Integração de Dados: realizada com a necessidade de ferramenta que permita sua orquestração, como: Hadoop, Apaches, EMR da Amazon, Mongo DB, entre outros.
9. Nuvem: suas soluções possuem grande potencial e infinitas áreas de uso, sendo IoT um dos mercados que mais se utilizam desse conceito, com o objetivo de utilizar somente recursos necessários, auxilia também à redução de custos, ampliação de faixa de mercado, precisão e escalonamento de aplicações que gerenciam grandes volumes de dados, dentre outros.
10. Self-service de dados: qualquer tecnologia que auxilie a simplificação dos processos de transformação de dados. Auxilia na capacitação de equipes de negócio e tomadores de decisão com a simplificação dos dados disponíveis para uso.

2.1.2 Etapas do Big Data

Para um dado ou um conjunto de dados brutos se transformarem em uma informação, para ser utilizada numa tomada de decisão, ele precisa passar por diversas etapas, podendo essas soluções, serem embarcadas ou distribuídas, e divididas em (TOTVS, 2018):

- Coleta - inserção dos dados para armazenamento, seguindo os formatos alinhados aos objetivos da empresa.
- Limpeza de dados - um pré-processamento é realizado a partir de métodos estatísticos, removendo incongruências e discrepâncias que podem ser fruto de erros de inserção.
- Mineração de dados - um processamento para a identificação de padrões, através de soluções computacionais que através de um olhar geral das informações, capta as mínimas mudanças nos dados.
- Análise de conteúdo - apresentar os dados, identifica padrões ou apresenta uma cadeia de eventos a partir de um acontecimento, classificando a análise descritiva, prescritiva e diagnóstica, respectivamente.
- Visualização de informações - etapa de remodelagem dos resultados das etapas anteriores, apresentando modelos que facilitam a interpretação dos resultados.

- Integração de dados - utilização dos dados de forma sistêmica e inteligente, sendo o Big Data Analysis uma medida a ser adotada como modelo de negócio, para isso, é importante que os dados estejam em convergência.

2.1.3 Data Mining

O termo “mineração de dados”, remete ao conceito de exploração de minérios, na qual os minérios são os dados, a mina são os bancos de dados e as ferramentas são os algoritmos de processamento desses dados (Mariano et al., 2021).

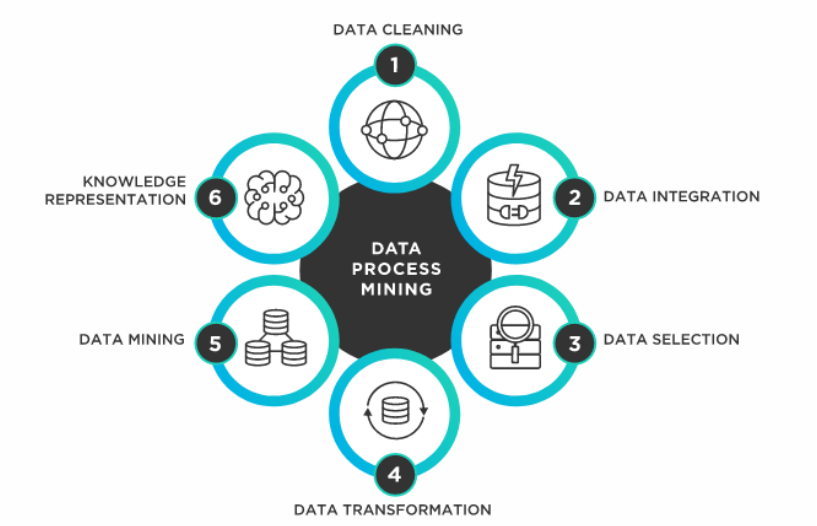

Figura 4 - Data Mining.
Fonte: www.tibco.com/pt-br/reference-center/what-is-data-mining

Dentre os aspectos de Data Analytics, a análise de dados possui ferramentas com diferentes focos: a Data Mining define um método de processamento de dados em grandes bases de dados com o objetivo de tentar descobrir uma relação entre as informações que podem ser valiosas para os negócios, de modo que, possam ser tomadas decisões eficientes diante dos resultados obtidos; o Big Data é o armazenamento desses dados, antes e depois do processamento; Business Inteligence (BI) tem como foco a coleta, organização, transformação e disponibilização dos dados já identificados como valioso para tomada de decisões (Souza, 2019).

O processo de Data Mining abrange 3 técnicas: estatística, base também de outras tecnologias; inteligência artificial, que busca se assemelhar ao pensamento humano para a resolução de problemas; e o machine learning, que consiste na união das duas técnicas anteriores, tendo como foco o desenvolvimento de algoritmos que aprendem com uma base de

dados, reconhecendo padrões e auxiliando na tomada de decisão. Dentre suas ramificações, podemos citar (Batista, 2016):

- Redes neurais - baseado na computação de ligações.
- Indução de regras - que consiste na detecção de padrões ou regras de um conjunto de dados.
- Árvores de decisão - que analisa o comportamento de entrada e saída de dados através de testes, a fim de, identificar quais possuem mais relevância para os valores finais do exame.
- Visualização - mapeia os dados minerados com base num filtro especificado, cuidando apenas da manipulação estatística, cabendo ao usuário a interpretação de comportamento dos dados.

2.1.4 Data Science

A ciência de dados surgiu como uma alternativa a necessidade de lidar com um grande conjunto de dados e informações em conjunto com algum aparato tecnológico. Ela abrange diversas disciplinas, como podemos ver na figura 5, se associando a ciência da computação, já que coleta, armazena e processa os dados; com estatística e matemática, a partir da análise e filtragem de dados; design gráfico, através da visualização dos dados; e especialização científica, com o questionamento, e refinamento dos dados visualizados (Morais et. al., 2018).

Figura 5 - Ciência de dados, interdisciplinar.
Fonte: Morais et. al., 2018.

Apesar de ser semelhante a outros termos da área como Inteligência Artificial e Machine Learning, cada uma tem suas características e aplicações, como (ORACLE, 2022):

- Inteligência Artificial - simular uma inteligência humana numa máquina.
- Ciência de Dados - subconjunto da IA, se referindo a áreas sobrepostas de estatísticas, método científico e análise de dados, usadas para extrair significado e percepções dos dados.
- Machine Learning - subconjunto da IA, técnicas para o descobrimento de padrões a partir de dados e fornecer aplicativos de IA.
    - Deep Learning - que permite resolver problemas mais complexos.

2.2 Ecossistema Hadoop

Sendo um software desenvolvido pela Apache Software Foundation em linguagem Java, ele contribuiu para o avanço da leitura de dados, proporcionando o processamento de um grande volume de dados com foco no desempenho e na eficiência. Sendo utilizado normalmente no ramo de computação distribuída, com a utilização de clusters, o Hadoop tinha como objetivo expandir um servidor para diversas máquinas, que viriam a oferecer mais computação e armazenamento. Devido a sua alta escalabilidade, é comumente utilizado no processamento de trabalhos em big data, que com o aumento de servidores, é possível aumentar o seu desempenho, conta também com um alto nível de disponibilidade e durabilidade, até mesmo durante o processamento de cargas de trabalho diversas ao mesmo tempo. Por possuir a combinação de durabilidade, escalabilidade e disponibilidade, é a ferramenta perfeita para o processamento de um grande volume de dados (EVEO, 2022).

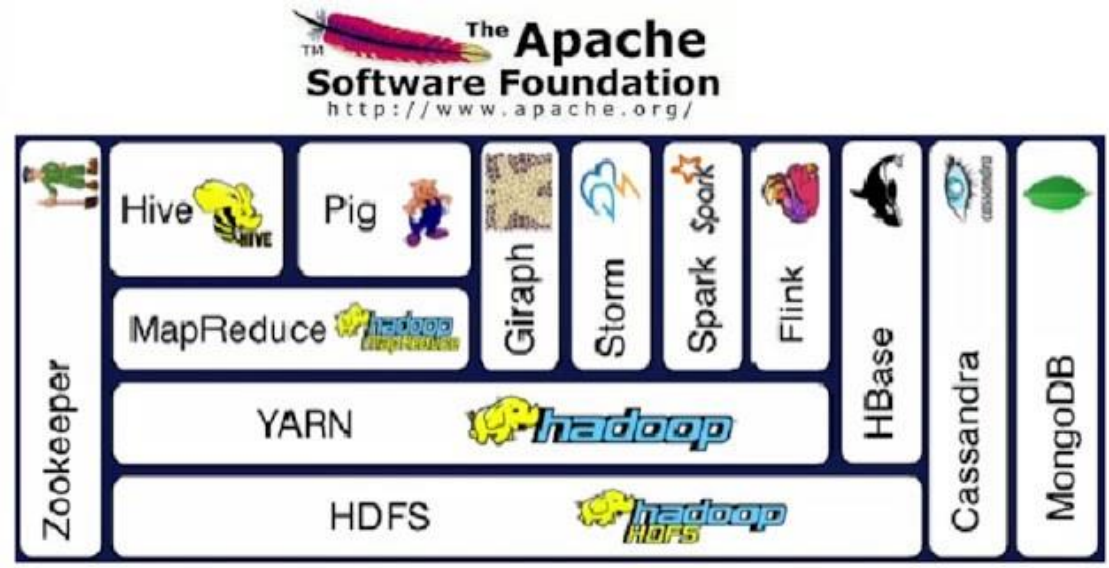

Figura 6 - Ecossistema Hadoop.
Fonte: joseantonio11.medium.com/.

O Hadoop é composto por módulos, cada um realiza uma tarefa específica para a sua execução em sistemas programados para realizar a análise de dados, como (EVEO, 2022):

- Distribuição de Sistemas de Arquivo - permite que os dados sejam armazenados de forma simples e disponível para uso entre diversos dispositivos de armazenamento vinculados.
- Hadoop Common - módulo que fornece ferramentas em java, com intuito de possibilitar a leitura dos dados armazenados no sistema de arquivo.
- Yarn - módulo responsável por administrar os recursos dos sistemas e por executar a análise.
- MapReduce - formata os dados de um banco de dados de modo adequado para análise, garantindo ferramentas que facilitam a exploração de dados.

O avanço da tecnologia se deu por novas tecnologias de armazenamento, novos meios de geração de dados, e com isso a capacidade de dispositivos e equipamentos de armazenamento teve um grande salto, porém, a velocidade de leitura e escrita não cresceu na mesma medida. Com isso o MapReduce se apresenta como uma solução para esse problema, trazendo a leitura e escrita em paralelo e com o uso de diversos discos, de modo a realizar o armazenamento sem perder velocidade, pois, ao particionar os dados em diversos dispositivos, sua leitura e escrita seria aumentada pela quantidade de dispositivos, aumentando o desempenho do processamento desses dados (EVEO, 2022).

Porém pode gerar dois problemas: o aumento da taxa de leitura e escrita pode ocasionar também no aumento da taxa de perda de dados, que pode ser evitada através de cópias de segurança em dispositivos diferentes; e por conta das partes dos dados que estão em discos diferentes no momento de análise dos dados, sendo diminuído através do método de "desembaralhar" os dados espalhados, através do processo de combinação de chaves e valores que pode encontrar em discos diferentes. O MapReduce é fácil e simples de utilizar, o desenvolvedor não lida com a parte teórica que inclui o grande processamento dos dados e de sistemas de arquivos distribuídos, e não precisa ter conhecimento de como funciona o processamento paralelo nem com o escalonamento de tarefas, sendo ambas funções do Hadoop (EVEO, 2022).

Sendo utilizada por grandes empresas, como podemos ver na figura 7.

Figura 7 - Empresas usando Hadoop.
Fonte: Kiran, 2021.

2.3 Docker

Docker consiste em uma plataforma aberta utilizada para desenvolver e rodar aplicações através de contêineres, é usada também para referir a empresa que a desenvolveu, a Docker.inc. A sua utilização é realizada sem a necessidade de qualquer acesso privilegiado a infraestrutura corporativa, agilizando o processo de desenvolvimento e manutenção de um serviço (Monteiro, 2021).

Por conter tudo necessário para sua execução, os contêineres se comunicam individualmente com o kernel do Sistema Operacional, com isso, os aplicativos de contêineres diferentes funcionam da mesma forma, através de uma arquitetura simplificada, como mostra a figura 8 (Wainstein, 2018).

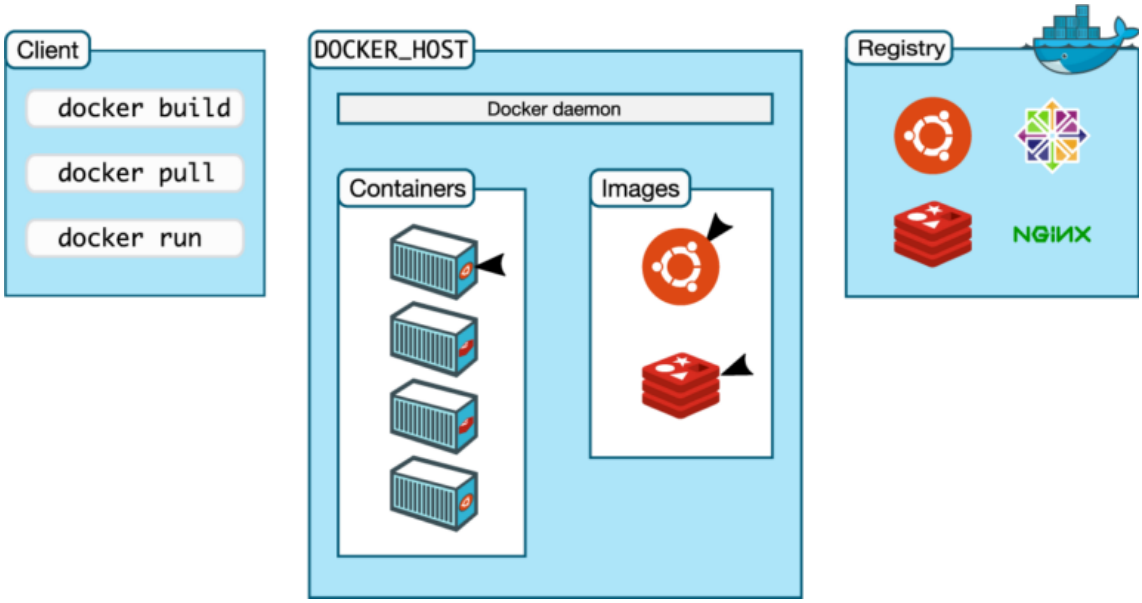

Figura 8 - Arquitetura Docker.
Fonte: Doerrfeld, 2021.

Viabilizou uma “linguagem” comum entre desenvolvedores e administradores de servidores, é utilizada para construir arquivos com as definições necessárias de infraestrutura e aspectos mais técnicos como, porta do serviço, dados de volumes externos bem como outras necessidades, Gomes (2020).

Diferente de softwares que com o tempo se tornam ineficientes por seu código-fonte em blocos, a conteinerização resolve esse problema passando por diferentes etapas: os micros serviços, desagregação do aplicativo em componentes menores; a adoção dos desenvolvedores a uma arquitetura para aumentar a eficiência operacional. O código-fonte é destinado para cada componente de aplicação, passando por vários estágios até o de produção, onde os contêineres garantem uma performance consistente através do encapsulamento da aplicação (Positivo, 2017).

É um sistema desenvolvido na linguagem de programação Go, que usa o Linux Container (LXC) como backend, fornecendo um ambiente virtual semelhante ao chroot (um utilitário de mudança de diretório com conceito de jaula), com um isolamento maior, o qual permite limitar recursos por contêiner (Positivo, 2017).

Sem a carga extra de um hypervisor e sem a necessidade de possuir uma cópia completa do sistema operacional como as máquinas virtuais, o contêiner compartilha as bibliotecas e o kernel com o seu hospedeiro, possibilitando que múltiplos serviços e aplicações sejam executadas isoladamente no mesmo host, como mostra a figura 9 (Gomes, 2020).

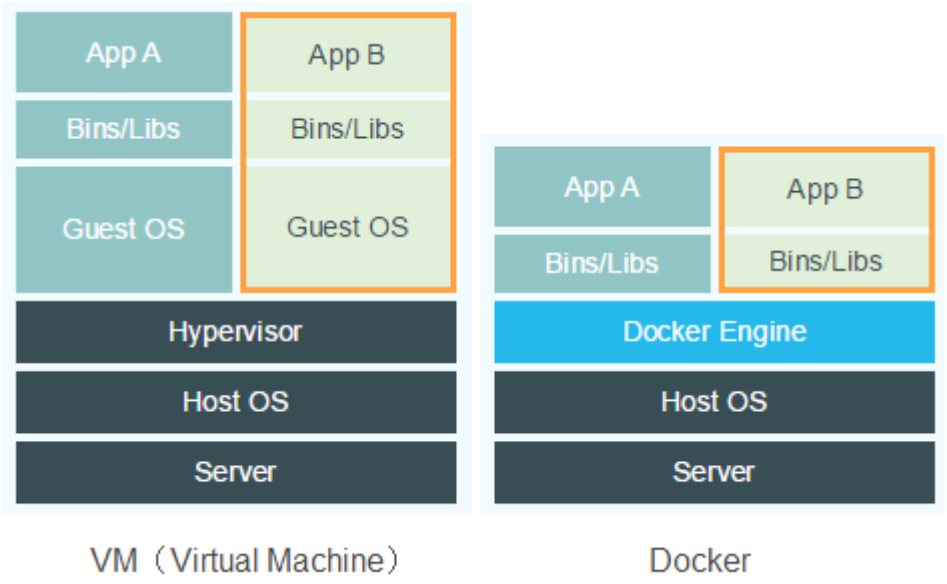

Figura 9 - Infraestrutura Docker.
Fonte: Gomes (2020).

As máquinas virtuais trabalham com uma imagem completa do sistema operacional, onde a execução do software está acima do servidor físico com a finalidade de emular determinado sistema de hardware, através do hypervisor, que virtualiza o servidor, cada VM conta com seu sistema operacional exclusivo, com seu próprio kernel, aplicativos, bibliotecas e variáveis, ocupando muito espaço e recursos. Nos primórdios dessa tecnologia foi possível alcançar a redução de custos e a possibilidade de recuperação de desastres, viabilizou a testagem de programas e o uso de servidores liberados para montar esses ambientes, uma grande evolução e novidade para a tecnologia existente na época, mas que possuía seus pontos negativos, e com a chegada do docker esses pontos negativos foram supridos, caracterizando um método de virtualização mais eficiente (Gomes, 2020).

Com o objetivo de gerir os recursos da máquina de forma eficiente para cada aplicação, o Docker utiliza de funcionalidades como *namespaces*, para a criação de ambientes isolados entre contêineres, bloqueando o acesso a recursos de outras aplicações, exceto quando declarado; e *cgroups*, o qual limita o uso de recursos de hardware disponível para cada aplicação, evitando com que uma aplicação consuma em excesso e interfira no funcionamento de outras. É possível criar uma rede de contêineres para que eles se conectem, tendo como padrão a *bridge*, como mostra na figura 10 (Resende et. al., 2020).

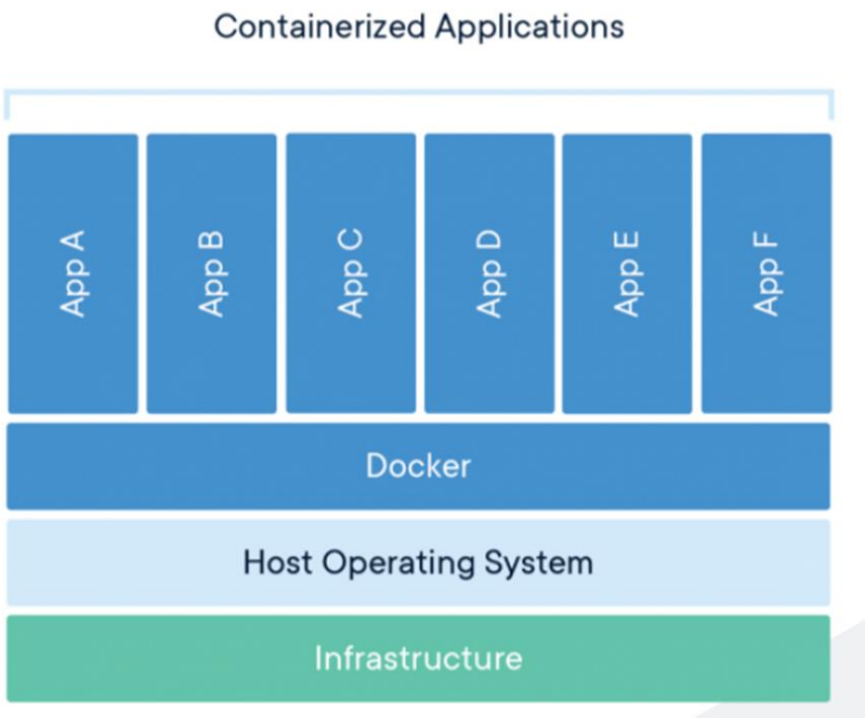

Figura 10 – Estrutura de contêineres e infraestrutura Docker.
Fonte: Resende et. al. (2020).

Uma empresa pode utilizar mais de uma ferramenta para suas soluções em big data, de modo a maximizar a eficiência de suas operações adequando-as a ferramenta que melhor atende as necessidades, contudo, ao se utilizar outras ferramentas ao mesmo tempo pode causar incompatibilidades, devido as diferentes soluções que podem estar utilizando um mesmo dado, e para esses casos que se é necessário que cada ferramenta tenha seu próprio espaço independente das disponibilidades de outras ferramentas, com isso, soluções em Docker resolvem os problemas, podendo (Wainstein, 2018):

- **Isolar ferramentas:** Ao assegurar um contêiner exclusivo para cada aplicação, com seus respectivos recursos para execução.
- **Operações de Análise agendadas**: são um tipo de processamento automatizado de dados que podem executar através de uma agenda ou horário específico, auxiliando quando grande volume de dados são enviados a uma organização em alta velocidade, que necessita de uma automação para manter a velocidade das tarefas. Contêineres do Docker adicionam a conveniência de processos programados, ao agendá-los sem a necessidade de uma interferência manual cada vez que for necessária.
- **Previsão de aplicações em ambientes diferentes**: possibilita o teste de diferentes tecnologias e ferramentas em diversos ecossistemas de big data, viabilizando a

interação de cliente-servidor e a respectiva visualização e processamento dos dados, sem a necessidade de outra máquina.

- **Construção de uma arquitetura de micro serviço**: facilita a transição do ambiente para a construção de micro serviços para aplicações em big data. Micro serviços são independentes e modulares e encaixam perfeitamente num ambiente de contêineres, que provê uma plataforma natural de implementação para essas soluções, gerando uma boa escalabilidade e qualidade de dados, que consequentemente gera uma segurança ao serviço desenvolvido.
- **Desenvolvimento de um sistema multi-nuvem de processamento de big data:** o sistema que processa os dados, bem como as suas ferramentas complexas de análise são desenvolvidos na nuvem, desse modo a extração, o processamento e a análise estão sendo feitos em nuvem, possibilitando assim, grandes insights de big data e diminuindo o custo ao terceirizar as máquinas.

Ao se utilizar Docker em ciência de dados, garante que todos trabalhem no mesmo ambiente, com as bibliotecas, serviços e as demais estruturas sem que precise instalá-las manualmente. É extremamente necessário visto que uma equipe de cientistas pode trabalhar com ubuntu, outra equipe com Windows e outra com MacOS, e não haverá o impeditivo para implementação e execução de uma aplicação, podendo ser replicado e distribuído para diversas máquinas ou filiais. Com essa solução em mente, e a partir do incentivo de compartilhamento de imagens e templates de sistemas iniciados pela Docker, tanto usuários como empresas compartilham modelos de ambientes para adequação e customização, dentre eles, empresas de ciência de dados, como a DSBrigade, que disponibilizam os arquivos que utilizam como base para qualquer projeto de Data Science, além de proverem artigos para auxiliar o uso dessas ferramentas (Resende et. al., 2020).

Conforme o docker foi se consolidando, as empresas viram um grande potencial e passaram a utilizar da tecnologia em suas aplicações e serviços, aumentando a demanda pela tecnologia até o momento em que grandes empresas como Google e Amazon, necessitam e proveem serviços em docker. Segundo Positivo (2017), podemos citar como benefícios:

- Economia de Recursos – por conta do compartilhamento de bibliotecas e kernel, a execução e um ou múltiplos contêineres, compartilham as mesmas bibliotecas, tendo uma economia de recursos sobre o disco. Caso senha necessário utilizar cinco contêineres, não terá um consumo de cinco vezes mais recursos, visto que seriam compartilhados.

- Disponibilidade maior do sistema – com a virtualização de sistema do docker sendo diferente do método de máquinas virtuais, há o compartilhamento de sistema operacional e outros componentes, deixando os processos mais ágeis e oferecendo disponibilidade maior.
- Possibilidade de Compartilhamento – os contêineres podem ser compartilhados para outros, ou até mesmo seus arquivos ou volumes, indicado quando se deseja ter persistência de dados e quando não se vincula ao host do contêiner; através do *namespace* o disco, RAM, processamento e acesso a rede são isolados, havendo o compartilhamento do kernel; compartilhamento em nuvem, por conta de uma biblioteca de imagens e ambientes prontos, podendo extrapolar o limite técnico, passando para questões de gerência, processo, manutenção e update.
- Gerenciamento facilitado – as clusters, que são grupos de máquinas físicas ou virtuais que executam os contêineres, precisam ser monitoradas constantemente, por conta disso, foram desenvolvidas ferramentas para essa função, como kubernetes e openshift. Possibilitando a execução de contêineres, essas ferramentas gerem também os sistemas de arquivos, e geram uma abstração que inclui um ou mais contêineres, armazenamento compartilhado e alternativas de operação.
- Similaridade dos Ambientes – a aplicação é transformada em uma imagem docker, permitindo que ela seja instanciada em diferentes plataformas, garantindo sua utilização tanto no ambiente de testes quanto no servidor de produção. Ambientes semelhantes impactam na análise de erros e a confiabilidade, podendo possuir um artefato de transição, como um docker com as dependências necessárias para execução do código compilado ou dinâmico.
- Aplicação como pacote completo – as imagens do docker consiste no empacotamento da aplicação e suas dependências, simplificando sua distribuição, bastando disponibilizar a imagem e permitir o acesso. Facilita a modificação e o compartilhamento da mesma, onde somente a alteração é transferida, tornando o desenvolvimento em diferentes ambientes um processo mais ágil e simplificado, além das imagens poderem ser armazenadas com tags, facilitando o versionamento de aplicações e serviços, que em caso de problema, é possível realizar o retorno para uma imagem com a tag anterior.
- Padronização e replicação – imagens docker são construídas por meio de arquivos de definição, assegurando a adoção de um determinado padrão, elevando a

confiabilidade na replicação, é possível realizar a mudança de um ou mais parâmetros de um arquivo de definição, facilitando a mudança de infraestrutura.

- Possibilidade de acessar a comunidade – a Docker disponibiliza um repositório de imagens docker, facilitando o acesso a modelos de infraestrutura de aplicações e serviços prontos para integrações. Sendo necessário o uso desse recurso, é possível utilizar as imagens do repositório e configurar parâmetros de adequação, visto que, as imagens oficiais costumam ser condizentes com as boas práticas.

A popularização do docker tem aumentado devido a esses benefícios que ela proporciona, sendo utilizada por grandes empresas de tecnologia como: Google, Amazon, Uber e Spotify, principalmente devido a sua eficiência de custo, manutenção, uso e desenvolvimento em comparação ao modelo de máquinas virtuais além de que diversas soluções de hospedagem adotaram essa tecnologia (Wainstein, 2018).

Sendo utilizada por grandes empresas, como podemos ver na figura 11.

Figura 11 - Empresas usando Docker.
Fonte: Kiran, 2021.

2.4 Kubernetes

O Kubernetes ou K8s, é uma ferramenta de código aberto robusta e sofisticada desenvolvida em linguagem Go, utilizado para o gerenciamento, automatização e dimensionamento de aplicações em contêineres ou micro serviços agrupados em unidades lógicas, ou nós de uma cluster, visando a facilitação de suas operações. Sua eficiência e eficácia, bem como um de seus grandes atrativos se dá por ter sido criada pela Google, trazendo consigo anos de experiência e aprimoramento de uma comunidade de desenvolvedores experientes. Teve sua abreviação para K8s devido a quantidade de letras entre o “K” e o “s” em Kubernetes (Doerrfeld, 2021).

Tem como objetivo esconder a complexidade de manusear inúmeros contêineres através de uma interface para as funcionalidades necessárias. Uma grande característica está em sua flexibilidade, pois ele disponibiliza aplicações que rodam de forma contínua e simples, independente da complexidade, é também uma ferramenta de código aberto, que torna possível

ser utilizado em cluster local, em sistemas de arquitetura híbrida e em sistemas de computação em nuvem, facilitando sua migração para diversos sistemas. É bem ampliável e automatiza muitas operações habituais, vindo da necessidade de empresas que implantam contêineres com frequência e que precisam administrá-los, auxiliando também, na descoberta de serviços, equilíbrio de carga, tem como características (Monteiro et. al., 2021):

- Escalabilidade - fornecendo dimensionamento horizontal com base na utilização de recursos.
- Alta Disponibilidade - possui uma camada de armazenamento confiável, oferecendo suporte a diversos back-ends de armazenamento distribuído, como AWS EBS (Amazon Elastic Block Store), Azure Disk, Google Persistent Disk, entre outros.
- Segurança - implementa princípios de segurança em diversos níveis: cluster, aplicativo e rede.
- Portabilidade - os clusters podem executar em qualquer sistema operacional, em diversas arquiteturas de processador, em sistemas de nuvem e sistemas diferentes de contêineres, como docker, podem ser adicionados.

Segue uma arquitetura cliente-servidor, os servidores mestre e trabalhador são constituídos de diversos componentes, como podemos ver na figura 12, com isso temos (Aquasec, 2022):

- Mestre
  - Cluster etcd - um armazenamento de valor simples e distribuído, utilizado para armazenar dos dados do cluster Kubernetes, objetos de API e detalhes de descoberta de serviço. Permite notificações ao cluster acerca de atualização de informações no armazenamento.
  - Kube API Server - o pilar principal de gerenciamento, recebe todas as solicitações para modificações, utilizada como interface do cluster. Por ser o único componente que armazena os dados, garante que os dados sejam armazenados no cluster etcd e que estejam de acordo com os padrões.
  - Kube Controller Manager - executa vários processos de controle diferentes em segundo plano, de modo a regular o estado da cluster e realizar tarefas de rotina.
  - Kube Scheduler - efetua o agendamento dos grupos de contêineres que a aplicação utiliza, de modo a utilizar os recursos eficientemente, agendando a aplicação ao nó que possui os recursos necessários para sua execução.

- Trabalhador
  - Kubelet - serviço principal, recebe regularmente as especificações novas ou modificadas através da Kube API Server, e garante que aos grupos e seus contêineres estejam funcionando corretamente.
  - Kube Proxy - serviço para lidar com sub-redes de host individual e disponibilizar os serviços ao acesso externo, executa o encaminhamento das solicitações para os respectivos grupos/contêineres nas diversas redes do cluster.

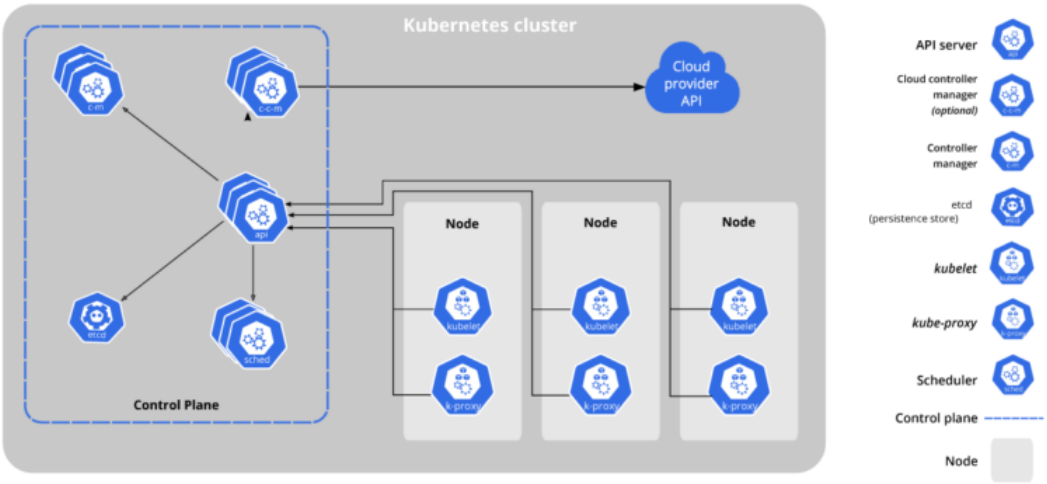

Figura 12 - Cluster Kubernetes.
Fonte: Doerrfeld, 2021.

Kubernetes foi desenvolvido inicialmente como uma ferramenta interna chamada Borg na Google, mas se tornou código aberto desde 2014, sendo uma das tecnologias mais utilizadas de contêineres, superando Apache Mesos, Docker Swarm e Nomad. Foi graduado como um projeto da CNCF (Cloud Native Computing Foundation) desde 2018 (adaptado Aquasec, 2022).

Por ter sido desenvolvido por uma empresa que executa uma quantidade massiva de contêineres por semana, um ponto negativo é ter sido construído para o uso pesado que uma empresa grande necessita, sendo um exagero para projetos pequenos, podendo ser substituído por Docker. Além de requerer treinamento inicial significativo, além de ser difícil manter e atualizar de tempos em tempos. Pode ser uma tecnologia de contêineres possui semelhanças e diferenças entre outras tecnologias, como o docker, na qual ambos são tecnologias abertas de nuvem, onde seus componentes são suportados em diversos serviços de computação em nuvem. A diferença é que o Docker encapsula aplicativos em contêineres em uma única máquina e o Kubernetes deve executá-los em uma cluster, podendo ser utilizados em conjunto. A empresa Booking.com construiu 500 aplicativos na plataforma em oito meses através do Kubernetes (Doerrfeld, 2021).

Sendo utilizada por grandes empresas, como podemos ver na figura 13.

Figura 13 - Empresas usando Kubernetes.
Fonte: Kiran, 2021.

2.5 Data Lake

Data lake consiste em um repositório central para armazenamento de uma grande quantidade de dados em seu formato original, estruturado ou não, utilizado em diversas soluções de big data (figura 14). Foi desenvolvido devido as limitações de repositórios de dados: caros e que não podiam utilizar os formatos modernos que as empresas necessitam processar, a necessidade de uma forma específica de armazenamento. Os dados podem ser armazenados em qualquer momento do estágio de processamento, sejam eles dados brutos que podem ser extraídos e armazenados em bancos de dados estruturados, bem como dados intermediários não tratados (vídeo, áudio, documento, dados de sensores) que são necessários para casos de uso de machine learning e outros tipos de análise complexa de dados (Databricks, 2022).

Figura 14 - Data Lake.
Fonte: Ribeiro, 2020.

Os data lakes são configurados em um cluster de hardware barato, permitindo que dos dados sejam colocados no "lago" caso necessite deles em outro momento, sem preocupação com a capacidade de armazenamento, podendo ser em cluster ou na nuvem, sendo economicamente viável a armazenar terabytes e até pentabytes de dados (Qubole, 2022).

Funciona num princípio de *schema-on-read*, onde não necessita de um esquema predeterminado de dados para o armazenamento, assim, só quando os dados são lidos para processamento que são analisados e adaptados para o esquema especificado, economizando tempo que seria gasto durante a definição de um esquema, permitindo o armazenamento dos dados em qualquer formato. Possibilita o acesso, preparo e análise dos dados mais rápido e de modo mais preciso, por armazenar dados em formatos não tradicionais, possibilita aos profissionais em análise, o acesso aos dados para diferentes casos de uso, como detecção de fraude (Talend, 2022).

A arquitetura da nuvem dos Data Lakes leva a seis vantagens para cargas de trabalho de big data, como (Qubole, 2022):

- Adaptabilidade - se adapta as mudanças nas cargas de trabalho e nas necessidades de negócio, a elasticidade possibilita uma eficiência no gerenciamento de dados.
- Agilidade - apesar das soluções levarem em torno de oito meses para implementação, é possível consultar os dados em questão de dias, agilizando o tempo das equipes de desenvolvimento, que podem se concentrar em outros aspectos da produção. Com a ajuda da nuvem, é possível ajustar e otimizar a configuração da máquina ou tamanho do cluster.
- Redução de custo - cargas de trabalho de big data consomem muito recurso computacional, com a arquitetura em nuvem, pode ser dimensionada de acordo com as necessidades.
- Segurança empresarial - soluções em nuvem oferecem diversos recursos de segurança, além de ser capaz de adotar as melhores práticas mais rapidamente.
- Alcance geográfico - permite escolher onde serão armazenados os dados, comumente usados para fins legais.
- Tolerância a falhas - Por ser um ambiente descentralizado, permite a recuperação de modo rápido em caso de desastre, possibilitando o uso de um outro nó em caso de falha, diminuindo assim, o tempo de manutenção e os possíveis gastos.

Para o processamento de dados de big data além do componente de armazenamento é necessário também o componente de computação, ambos podem estar localizados tanto localmente quanto em nuvem, se adequando as necessidades do projeto, podendo se vincular a outras tecnologias como (Talend, 2022):

Hadoop - pode utilizar plataformas de data lake como Cloudera e HortonWorks, e ferramentas de ingestão, preparação e extração de dados como Hive, Pig, Flume, entre outros.

- Mais familiaridade com tecnólogos
- Barato, por ser código aberto
- Diversas ferramentas disponíveis para integração com Hadoop
- Fácil de dimensionar
- A localização dos dados agiliza a computação

AWS - oferece diversos produtos para soluções de data lake. Amazon Simple Storage Service (Amazon S3) para a função de armazenamento, e Snowball, Direct Connect e Kinesis (Streams e Firehose) para a ingestão de dados.

- Grande conjunto de produtos, rico em recursos
- Flexibilidade de produto de acordo com a necessidade
- Baixo custo
- Alta segurança e conformidade
- Divisão de computação e armazenamento, de forma a dimensionar de acordo com as necessidades.

Azure - data lake oferecido pela Microsoft, possui uma camada de armazenamento como Azure Data Lake Store (ADLS) e uma de análise, Azure Data Lake Analytics e HDInsight.

- Fácil gerenciamento, por ser em nuvem
- Serviço de análise com ótimas funcionalidades
- Fácil migração de Hadoop
- Fácil adaptação do Hadoop
- Integração com Active Directory segura

2.6 Programação em R

É uma linguagem de programação e um ambiente de software livre utilizado na computação estatística, análise de dados e pesquisa científica, é popularmente utilizada por analistas de dados pesquisadores e profissionais de marketing para processamento, análise e visualização de dados, com diversas vantagens como na figura 15. Atualmente o projeto R é mantido pela R Foundation e pelo R Core Team (Hackanons, 2020).

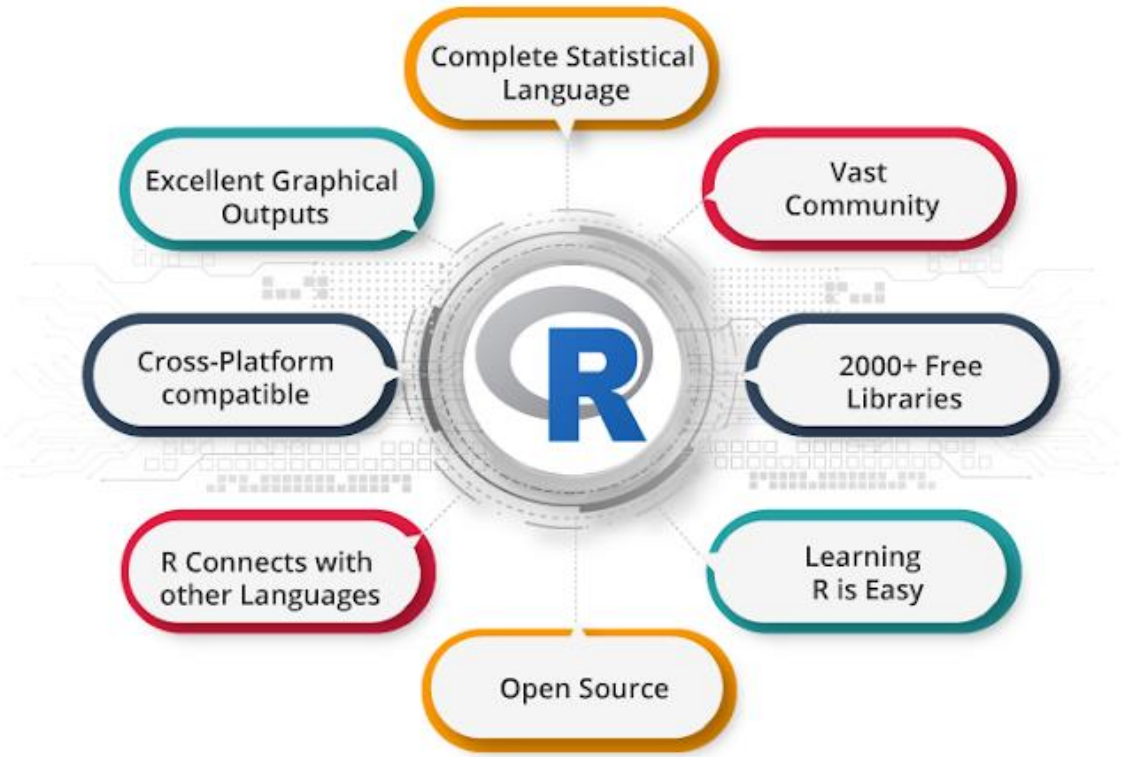

Figura 15 - Vantagens Programação em R.
Fonte: Hackanons, 2020.

Dentre suas características pode-se destacar (Lacerda et. al., 2021):

- Multiparadigma - permite o uso de programação estrutural e orientada a objetos.
- Linguagem de alto nível - não é necessário desenvolver funções simples, permitindo focar no desenvolvimento de aplicações de alto nível.
- Dinâmica - permite a converter tipos de objetos.
- Fracamente tipada - não necessita que os objetos sejam do mesmo tipo para realizar as operações

Por ter sido construído especificamente para análise estatística e visualização de dados não limita seu uso, de modo a ser amplamente utilizada como linguagem de programação. Cientistas, pesquisadores e estudantes utilizam para analisar os resultados de experimentos,

além de diversos negócios de ramos diferentes utilizarem para extrair padrões de uma grande quantidade de dados gerados diariamente, como (Miller, 2021):

- Fintech - empresas que lidam com serviços financeiros, para desenvolver modelos de risco de crédito e outras análises de risco, usado também para detecção de fraudes, modelagem de hipotecas e volatilidade, avaliação de clientes e simulações.
- Pesquisa - utilizada para pesquisas e projetos acadêmicos, por exemplo a Universidade da Califórnia que ensina estatística e análise de dados através da R.
- Varejo - usado para avaliação de risco e para desenvolver estratégias de marketing.
- Governo - usado na previsão de desastres e do clima, como recursos de visualização de dados para desenvolver imagens de previsão do tempo, bem como dados geográficos.
- Jornalismo de Dados - profissionais que utilizam os dados para contar histórias, extraindo insights sobre o mundo e como vivemos de dados públicos. Sejam para contar sobre a economia através de dados financeiros, ou sobre crimes através dos dados governamentais e policiais, com as ferramentas de visualização.
- Mídia Social - usada na análise para a segmentação de clientes em potencial, personalização de anúncios, utilizando um banco de dados geral com dados pessoais.
- Assistência médica - usado na genética, bioinformática, descoberta de drogas e epidemiologia.
- Fabricação - análise de feedback dos clientes, para o aperfeiçoamento dos produtos.

A linguagem R é uma das mais adequadas para manipulação de dados, análise, mineração, processamento e análise gráfica, segundo Hackanons, 2021:

- Capacidade de manipulação e armazenamento de dados.
- Diversas ferramentas para análise precisa.
- Ferramentas e recursos integrados para análise de big data.
- Facilidades gráficas para visualizar grandes massas de dados.
- Uma linguagem de programação elegante composta de loops, instruções, funções, recursos de entrada e saída e muito mais.

Sendo utilizada por grandes empresas, como podemos ver na figura 16.

Figura 16 - Empresas usando Programação em R.
Fonte: Kiran, 2021.

2.7 Outras tecnologias

Big data é um conceito extenso e amplamente utilizado, possuindo inúmeras ferramentas com características específicas, que de acordo com Kiran, 2021, podemos classificar:

- Armazenamento de dados
  - Mongo DB - alternativa ao esquema de banco de dados relacionais, que oferece uma ampla variedade de tipos de dados em arquiteturas distribuídas, atualmente utilizado pelas empresas da figura 17 (Kiran, 2021).

Figura 17 - Empresas usando Mongo DB.
Fonte: Kiran, 2021.

  - RainStor - Projetado pela empresa de mesmo nome, tem o objetivo de gerenciar e analisar big data, que através de técnicas de desduplicação organiza o processo de armazenamento de dados, atualmente utilizado pelas empresas da figura 18 (Kiran, 2021).

Figura 18 - Empresas usando RainStor.
Fonte: Kiran, 2021.

  - Hunk - Permite o acesso aos dados de clusters Hadoop remotos através de índices virtuais, utiliza a linguagem de processamento Splunk para relatar e visualizar dados de fontes Hadoop e NoSQL (Kiran, 2021).

- Mineração de Dados
  - Presto - mecanismo de consulta SQL distribuído de código aberto para realizar consultas analíticas interativas. Permite consultar dados em Hive, Cassandra, Bancos de Dados Relacionais e Armazenamentos de Dados Proprietário, atualmente utilizado pelas empresas da figura 19 (Kiran, 2021).

Figura 19 - Empresas usando Presto.
Fonte: Kiran, 2021.

  - RapidMiner - solução caracterizada pela sua interface gráfica de usuário poderosa e robusta, que permite criar, entregar, e manter análises preditivas, possui suporte a scripts em várias linguagens, atualmente utilizado pelas empresas da figura 20 (Kiran, 2021).

Figura 20 - Empresas usando RapidMiner.
Fonte: Kiran, 2021 (Kiran, 2021).

  - Elasticsearch - mecanismo de pesquisa de texto completo distribuído, com uma interface da Web HTTP e documentos JSON sem esquema, atualmente utilizado pelas empresas da figura 21 (Kiran, 2021).

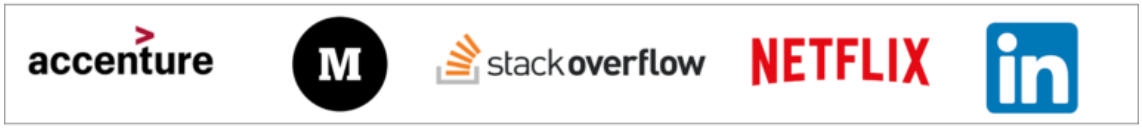

Figura 21 - Empresas usando Elasticsearch.
Fonte: Kiran, 2021.

- Análise de Dados
    - Kafka - plataforma de streaming distribuído, com três recursos: Editor, Assinante e Consumidor, semelhante a um sistema de mensagens corporativas, atualmente utilizado pelas empresas da figura 22 (Kiran, 2021).

Figura 22 - Empresas usando Kafka.
Fonte: Kiran, 2021.

    - Spark - fornece recursos de computação em memória, fornecendo velocidade, modelo de execução com suporte a diversos aplicativos e APIs Java, Scala e Python que facilitam o desenvolvimento, atualmente utilizado pelas empresas da figura 23 (Kiran, 2021).

Figura 23 - Empresas usando Spark.
Fonte: Kiran, 2021.

    - Blockchain - usado para serviços essenciais, como pagamento, depósito e título, reduz fraudes, aumenta privacidade financeira, acelera transações e agrupa os mercados internacionais. Tem como características (Kiran, 2021):

        - Razão Compartilhada: Sistema Distribuído de registros em uma rede de negócios.
        - Contrato Inteligente: Os termos comerciais são incorporados no banco de dados de transações e executados com transações.
        - Privacidade: transações seguras, autenticadas e verificáveis
        - Consenso: Todas as partes concordam com as transações verificadas na rede.

    Atualmente utilizado pelas empresas da figura 24.

Figura 24 - Empresas usando Blockchain.
Fonte: Kiran, 2021.

- Visualização de dados
  - Tableau - ferramenta de visualização de dados de BI poderosa e de crescimento rápido, análise rápida e visualizações em formato de painéis e planilhas, atualmente utilizado pelas empresas da figura 25 (Kiran, 2021).

Figura 25 - Empresas usando Tableau.
Fonte: Kiran, 2021.

- Tecnologias emergentes
  - TensorFlow - possui um ecossistema abrangente e flexível de ferramentas, bibliotecas e recursos da comunidade, permitindo criar e implantar aplicativos e impulsionar o aprendizado de máquina, atualmente utilizado pelas empresas da figura 26 (Kiran, 2021).

Figura 26 - Empresas usando TensorFlow.
Fonte: Kiran, 2021.

  - Apache Beam - fornece uma camada de API portátil para a construção de pipelines de processamento de dados paralelos, atualmente utilizado pelas empresas da figura 27 (Kiran, 2021).

Figura 27 - Empresas usando Beam.
Fonte: Kiran, 2021.

- Apache AirFlow - um sistema de automação e agendamento de fluxos de trabalho, para a criação e gerenciamento de pipeline de dados, fornecendo manutenção, teste e controle de versão mais fáceis, atualmente utilizado pelas empresas da figura 28 (Kiran, 2021).

Figura 28 - Empresas usando AirFlow.
Fonte: Kiran, 2021.

2.8 Tabela comparativa de tecnologias

Tecnologia	Categoria	Característica	Vantagem
Docker	Gerenciamento Contêineres	Isolamento de aplicações, multiplataforma, unificação de ambientes, implementável em outras tecnologias.	Baixo custo, disponível, escalável, gerenciamento facilitado.
Kubernetes	Gerenciamento	Gerenciamento, automatização e dimensionamento de contêineres, suporte a diversas tecnologias.	Eficiente, escalável, disponível, seguro e portátil.
Data Lake	Armazenamento	Tecnologia de armazenamento de dados distribuído, armazenamento em qualquer formato.	Adaptável, ágil, baixo custo, segurança empresarial, alcance geográfico e tolerante a falhas.
Hadoop	Armazenamento	Armazenamento e Processamento distribuído.	Durável, escalável e distribuído, disponível e de baixo custo.
Mongo DB	Armazenamento	Armazenamento em coleções com documentos diversos.	Performance, escalável e baixo custo.
RainStor	Armazenamento	Gerenciar e analisar dados para grandes empresas, técnicas de deduplicação de dados (dados únicos).	Elimina cópias redundantes, redução de estresse de armazenamento.
Presto	Mineração	Consulta em bancos de dados diversos, preparação leve de dados.	Processamento eficiente em grandes bancos de dados, leve organização dos dados.
RapidMiner	Mineração	Interface gráfica poderosa e robusta, suporte a scripts.	Simples e eficiente na criação de modelos preditivos.
ElasticSearch	Mineração	Diversas opções de busca, utiliza filtros para agilizar as buscas.	Rápido, escalável, eficiente e de fácil implementação.
Kafka	Análise	Plataforma de streaming de dados distribuída.	Eficiente, versátil, disponível, transmissão assíncrona.
Spark	Análise	Computação em memória, suporte a diversos aplicativos.	Escalável e distribuído, rápido com o processamento em memória.

Programação em R	Análise	Linguagem de alto nível, multiparadigma, fracamente tipada, dinâmica.	Alto nível de processamento e armazenamento, análise precisa, recursos gráficos.
Blockchain	Análise	Funções essenciais financeiras (Pagamentos, depósitos, títulos).	Redução de fraude, privacidade financeira, transações ágeis.
Tableau	Visualização	Processamento de dados, análise rápida e visualizações em painéis e planilhas.	Fácil implementação, visualização rápida.
TensorFlow	Desenvolvimento	Framework de desenvolvimento em aprendizado de máquina.	Grande conjunto de ferramentas, bibliotecas e recursos.
Beam	Processamento	Cria e executa conjuntos de instruções para o processamento em lote e streaming de dados.	Eficiente e suporta diversos runners (spark, google cloud dataflow, flink e samza).
AirFlow	Automação de fluxos	Automatiza e agenda fluxos de trabalho.	Fácil manutenção, controle de versão e teste.

3. CONSIDERAÇÕES FINAIS

Então, no presente trabalho são apresentadas as tecnologias de big data mais utilizadas para os principais segmentos de aplicação: armazenamento, tratamento e visualização dos dados, abordando de forma mais aprofundada as tecnologias amplamente utilizadas que não estão inseridas nesses segmentos como o conceito de conteinerização e lagos de dados, e o gerenciamento dessas tecnologias de forma integrada, em uma cluster, podendo ser automatizadas, com o objetivo de atingir um nível corporativo de eficiência e excelência.

Portanto, foi possível compreender, os aspectos relevantes de Big Data, os problemas que geraram seu desenvolvimento, sua evolução e os benefícios de sua utilização. E através das características e vantagens de cada tecnologia, como: simplicidade, implementação em setores variados, compatibilidade com outras ferramentas e baixo custo, por exemplo, é possível avaliar os motivos pelos quais são amplamente utilizadas, seja no âmbito comercial ou pessoal, e o potencial que possuem para a resolução de problemas sem a limitação de ramo de trabalho.

4. TRABALHOS FUTUROS

Como possíveis trabalhos que podem ser desenvolvidos a partir dessa pesquisa de análise de tecnologias, seria a utilização de mais de uma tecnologia para o desenvolvimento de uma solução para Data Science utilizando uma aplicação de gerenciamento de contêineres, como Hadoop, no desenvolvimento de aplicações e micro serviços para tornar menos custosa e mais eficiente o tratamento de dados, a fim de agilizar a tomada de decisão.

Outra sugestão seria a utilização de serviços de nuvem, para o desenvolvimento de uma aplicação que analisa, trata e processa dados disponibilizados pelo governo, para a identificação de padrões de doenças, se beneficia da capacidade computacional e da escalabilidade que a nuvem disponibiliza, para a redução de custos e auxiliar a tomada de decisões de natureza política.

Trabalho publicado no Github: https://github.com/NelsonRod/TCC

REFERÊNCIAS

Aquarela. **7 características importantes para diferenciar BI, Data mining e Big Data.** Disponível em: https://www.aquare.la/7-caracteristicas-importantes-para-diferenciar-bi-data-mining-e-big-data/ . Acesso em: 29 de mar. 2022.

Aquasec. Disponível em: https://www.aquasec.com/cloud-native-academy/kubernetes-101/kubernetes-complete-guide/ . Acesso em: 20 de maio de 2022.

Batista, Matheus (2016). **DESMISTIFICANDO O MUNDO DO BIG DATA.** Disponível em: https://cepein.femanet.com.br/BDigital/arqPics/1511420203P634.pdf . Acesso em: 12 de maio de 2022.

Cetax. **Big Data: O que é, conceito e definição.** Disponível em: https://www.cetax.com.br/blog/big-data/ . Acesso em: 15 de mar. 2022.

Databricks. **Introduction to Data Lakes.** Disponível em: https://databricks.com/discover/data-lakes/introduction. Acesso em: 20 de maio de 2022.

Doerrfeld, Bill (2021). **What's the Difference Between Docker and Kubernetes?**. Disponível em: https://containerjournal.com/editorial-calendar/best-of-2021/whats-the-difference-between-docker-and-kubernetes/#:~:text=The%20difference%20between%20the%20two,Kubernetes%20can%20be%20used%20independently. Acesso em: 20 de maio de 2022.

EVEO. **Como o Hadoop MapReduce se relaciona com Big Data?.** Disponível em: https://blog.eveo.com.br/hadoop-mapreduce-big-data . Acesso em: 6 de maio de 2022.

Gomes, E., & Braga, F. (2017). **Inteligência Competitiva Tempos Big Data**. Editora Alta Books. https://integrada.minhabiblioteca.com.br/books/9788550804101 .

Gomes, Rafael (2020). **Docker para Desenvolvedores**. Leanpub. Disponível em: https://stack.desenvolvedor.expert/appendix/docker/oquee.html . Acesso em: 6 de maio de 2022.

Hackanons (2020). **Is R programming language? – Things you should know.** Disponível em: https://hackanons.com/2020/11/is-r-programming-language.html Acesso em: 30 de maio de 2022.

Kiran, Ravi (2021). **Top Big Data Technologies that you Need to know**. Disponível em: https://www.edureka.co/blog/top-big-data-technologies/. Acesso em: 24 de maio de 2022.

Lacerda, P.S.P. D., Pereira, M. A., & Lenz, M. L. et al. (2021). **Programação em Big Data com R**. Grupo A. https://integrada.minhabiblioteca.com.br/books/9786556901091/.

Lorenzi, Larissa. **10 soluções de big data para uma análise de dados moderna.** Disponível em: https://blog.indicium.tech/10-solucoes-de-big-data-para-uma-analise-de-dados-moderna/ . Acesso em: 8 de mar. 2022.

Mariano, D.C. B., Marques, L. T., & Silva, M. S. et al. (2021). **Data Mining.** Grupo A. Disponível em: https://integrada.minhabiblioteca.com.br/books/9786556900292 .

Marquesone, Rosangela (2017). **Big Data: técnicas e tecnologias para extração de valor dos dados.** Editora Casa do Código.

Medeiros, Alan (2019). **Uma abordagem para o gerenciamento de infraestruturas virtualizadas em ambientes de internet das coisas.** Disponível em: http://dspace.bc.uepb.edu.br/jspui/bitstream/123456789/20451/1/PDF%20-%20Allan%20Medeiros%20de%20Lima.pdf .

Miller, Stephan (2021). **What is R used for?.** Disponível em: https://www.codecademy.com/resources/blog/what-is-r-used-for/. Acessado em: 30 de maio de 2022.

Monteiro, E. R., Cerqueira, M.V. B., & Serpa, M.D. S. et al. (2021). **DevOps.** Grupo A. https://integrada.minhabiblioteca.com.br/books/9786556901725

Morais, I.S. D., Gonçalves, P.D. F., & Ledur, C. L. et al. (2018). **Introdução a Big Data e Internet das Coisas (IoT)**. Grupo A. https://integrada.minhabiblioteca.com.br/books/9788595027640 .

ORACLE. **O que é Big Data?** Disponível em: https://www.oracle.com/br/big-data/what-is-big-data/. Acesso em: 15 de mar. 2022.

Positivo (2017). **Container docker: o que é e quais são as vantagens de usar?** Disponível em: https://www.meupositivo.com.br/panoramapositivo/container-docker/. Acesso em: 10 de maio de 2022.

Qubole (2020). **WHY ARE DATA LAKES IMPORTANT FOR BIG DATA?**. Disponível em: https://www.qubole.com/why-are-data-lakes-important-for-big-data. Acesso em: 29 de maio de 2022.

Resende, Antônio, Cardoso, Jon (2020). **Docker para projetos de ciência de dados: porque é importante e os principais conceitos.** DSBrigade. Disponível em: https://blog.dsbrigade.com/docker-para-ciencia-de-dados/ . Acesso em: 18 de abril de 2022.

Ribeiro, Janete (2020). **Data Lake e Delta Lake, você sabe a diferença?.** Disponível em: https://escoladeia.com.br/data-lake-e-delta-lake-voce-sabe-a-diferenca/ . Acesso em: 20 de maio de 2022.

Souza, Clayton (2019). **Business Intelligence X Big Data qual a diferença.** BDASolutions. Disponível em: https://bdasolutions.com.br/2019/06/business-intelligence-x-big-data-qual-a-diferenca/ . Acesso em: 29 de mar. 2022.

Talend (2022). **What is a Data Lake?**. Disponível em: https://www.talend.com/resources/what-is-data-lake/. Acesso em: 26 de maio de 2022.

TOTVS (2018). **Etapas de Big Data: como os dados viram insights.** Disponível em: https://www.totvs.com/blog/negocios/etapas-de-big-data/ . Acesso em: 23 de abril de 2022.

Wainstein, Limor (2018). **Docker use cases – How to handle big data with Docker.** Big Data Made Simple. Disponível em: https://bigdata-madesimple.com/docker-use-cases-how-to-handle-big-data-with-docker/ . Acesso em: 22 de abril de 2022.