# Frameworks de Big Data

UNIDADE 04

## Framework de big data: uma visão geral

*Nesta semana, vamos explorar os conceitos fundamentais no uso de soluções de frameworks para big data. Vamos destacar os principais desafios enfrentados devido ao volume, velocidade e variedade dos dados. É aqui que a importância dos frameworks especializados, como Hadoop e Spark, se destaca, lidando com essa complexidade na ingestão de dados. Além disso, vamos examinar brevemente estudos de caso e aplicações práticas. Ao final, faremos uma introdução ao Spark, entendendo essa ferramenta no contexto de big data.*

Hoje, vamos explorar o fascinante mundo do *big data* e como ele impulsiona as operações das grandes empresas, como a Netflix. Já parou para pensar em como a Netflix lida com a montanha de dados gerados diariamente por seus usuários? Aqui não se trata apenas de fornecer conteúdo, mas sim de personalizar recomendações com base nas preferências individuais de cada usuário. Como eles conseguem isso? A resposta está no uso eficiente do *big data*, que é o foco da nossa aula hoje.

### A importância do big data na Netflix

> *"A Hadoop se tornou o padrão de fato para gerenciar e processar centenas de terabytes a petabytes de dados. Na Netflix, nosso armazém de dados baseado em Hadoop é de escala petabyte e está crescendo rapidamente." (Fonte: Netflix Tech Blog)*

Essa abordagem demonstra claramente como o *big data* se tornou uma ferramenta essencial para entender e atender às demandas dos consumidores no mundo digital.

Agora, você deve estar se perguntando: como as empresas podem aproveitar ao máximo o potencial do *big data* para se destacarem em um mercado altamente competitivo?

## Conceitos fundamentais dos *frameworks* de *big data*

Definição de *big data* e sua relevância atual:

O *big data* refere-se ao **volume, variedade e velocidade** dos dados que inundam as empresas diariamente. São dados provenientes de diversas fontes, como transações comerciais, mídias sociais, dispositivos móveis, entre outros (Marquesone, 2020).

---

- Volume: faz referência à dimensão sem precedentes do volume de dados.
- Variedade: se refere à diversidade dos tipos e fontes de dados que estão sendo gerados e coletados.
- Velocidade: a velocidade com que os dados são coletados, analisados e utilizados

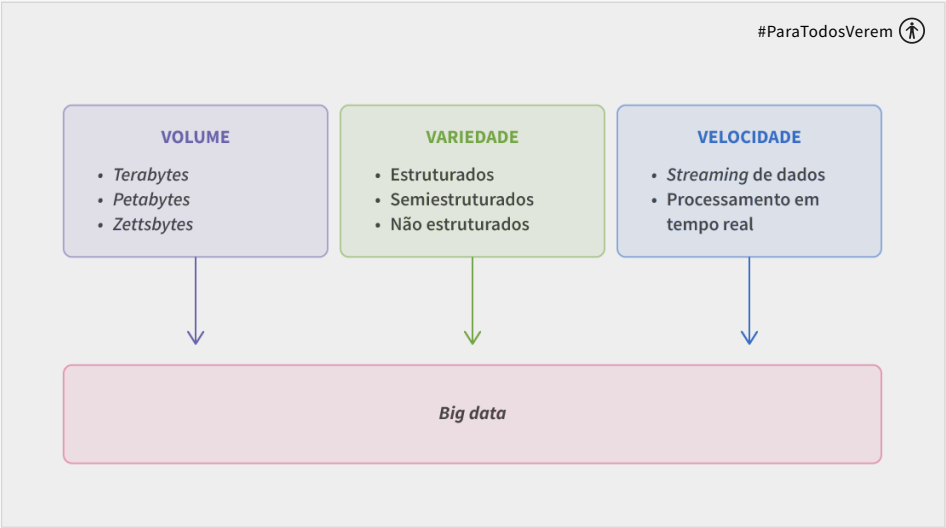


Figura 1: Os 3 Vs do *big data*. Fonte: Adaptado de Marquesone, 2020.

Agora que entendemos os fundamentos do *big data*, vamos mergulhar mais profundamente em um dos principais *frameworks* utilizados para lidar com essa imensidão de dados: o Hadoop

## Hadoop: Desvendando o Gigante do Big Data

O Hadoop é um dos principais *frameworks* de *big data*. Ele oferece uma plataforma completa para armazenar, processar e analisar grandes volumes de dados.

### O que é o Hadoop?

O Hadoop é um ***framework open source***, escrito em Java, que permite processar **grandes volumes de dados** de forma **eficiente e escalável**. Ele funciona como um maestro, orquestrando um conjunto de computadores comuns (*commodities*) para trabalhar em conjunto como um ***cluster* de alto desempenho**. Na figura 2, é mostrado o surgimento da ferramenta até o primeiro caso de uso real de aplicação

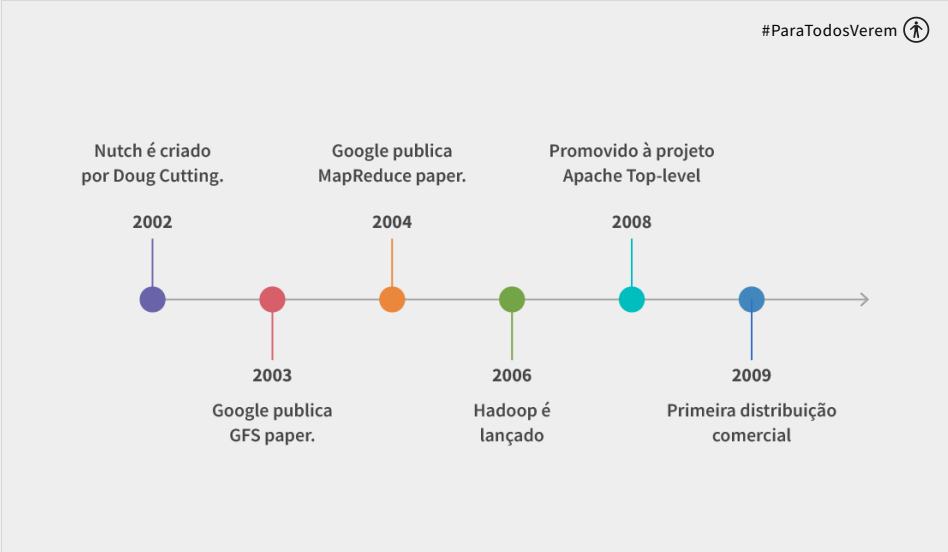


Figura 2: Evolução do Hadoop. Fonte: Adaptado de Santos, 2023.

**Mais que um armazém, uma plataforma:**

O Hadoop não se limita a armazenar dados. Ele oferece uma plataforma completa para:

- **Armazenar grandes volumes de dados** em seu sistema de arquivos distribuído, o **HDFS** (*Hadoop Distributed File System*).
- **Processar esses dados em paralelo** usando o modelo de programação **MapReduce**.
- **Analisar e extrair *insights*** valiosos dos dados com ferramentas como **Hive**, **Pig** e **Spark**.

## Ecossistema do Hadoop

O ecossistema do Hadoop é uma coleção de módulos interligados que constituem uma infraestrutura para computação distribuída. Desenvolvidos pela Apache Software Foundation, esses módulos são fundamentais para diversos projetos de código aberto. Os principais componentes incluem:

---

- **Hadoop Common**: conjunto de utilitários compartilhados que oferecem suporte a outros módulos do Hadoop.
- **HDFS**: Sistema de arquivos distribuído do Hadoop que fornece acesso aos dados da aplicação com alta taxa de transferência.
- **Hadoop Yarn**: *framework* para agendamento de tarefas e gerenciamento de recursos do cluster.
- **Hadoop MapReduce**: sistema baseado no Yarn para o processamento paralelo de grandes conjuntos de dados.
- **Hadoop Ozone**: sistema de armazenamento de objetos do Hadoop.
- **Hadoop Submarine**: motor de aprendizado de máquina do Hadoop

Apresentamos no quadro 1 outros projetos relacionados ao Hadoop da Apache que incluem:

| Ferramenta | Descrição | Funcionalidade principal |
|---|---|---|
| **Ambari** | Ferramenta *web* para suporte, gerenciamento e monitoramento de outros módulos do Hadoop. | Simplifica a administração do Hadoop, automatizando tarefas e fornecendo uma interface amigável. |
| **Avro** | Sistema de serialização de dados. | Permite a serialização eficiente de dados em um formato binário, facilitando a troca de dados entre diferentes sistemas. |
| **Cassandra** | Banco de dados distribuído tolerante a falhas e altamente escalável. | Ideal para armazenar grandes quantidades de dados que precisam ser acessíveis em tempo real, mesmo em caso de falhas. |
| **Chukwa** | Sistema para coleta de dados e monitoramento de sistemas distribuídos. | Coleta e monitora *logs* e métricas de sistemas Hadoop em tempo real, facilitando a identificação de problemas e a otimização do desempenho. |
| **HBase** | Banco de dados distribuído e escalável com suporte para armazenamento de dados e estruturas em grandes tabelas. | Ideal para armazenar grandes conjuntos de dados estruturados que precisam ser acessados de forma rápida e eficiente. |
| **Hive** | Infraestrutura de armazenamento de dados que oferece sumarização e consultas para fins específicos. | Facilita a análise de grandes conjuntos de dados armazenados no Hadoop, fornecendo uma linguagem SQL familiar para realizar consultas. |
| **Mahout** | Sistema para aplicativos de aprendizado de máquina e bibliotecas com funções de mineração de dados. | Permite a criação de aplicações de *machine learning* e *data mining* utilizando o Hadoop, oferecendo algoritmos e ferramentas para diversas tarefas. |
| **Pig** | Linguagem de consulta de alto nível orientada a fluxo de dados, com uma estrutura de execução para computação paralela. | Facilita o processamento de grandes conjuntos de dados em lote, utilizando uma linguagem de alto nível que abstrai a complexidade do Hadoop. |
| **Spark** | Motor de computação de propósito geral, rápido e altamente eficiente para trabalhar com dados, oferecendo um modelo de programação que suporta diversos tipos de aplicativos. | Pode ser usado para diversas tarefas de *big data*, como processamento em lote, *streaming*, *machine learning* e análise de grafos, oferecendo alto desempenho e escalabilidade. |

Juntos, eles formam o Hadoop, uma plataforma completa para domar o *big data* e extrair *insights* valiosos.  Uma representação visual do ecossistema do universo Hadoop pode ser vista na figura 3.

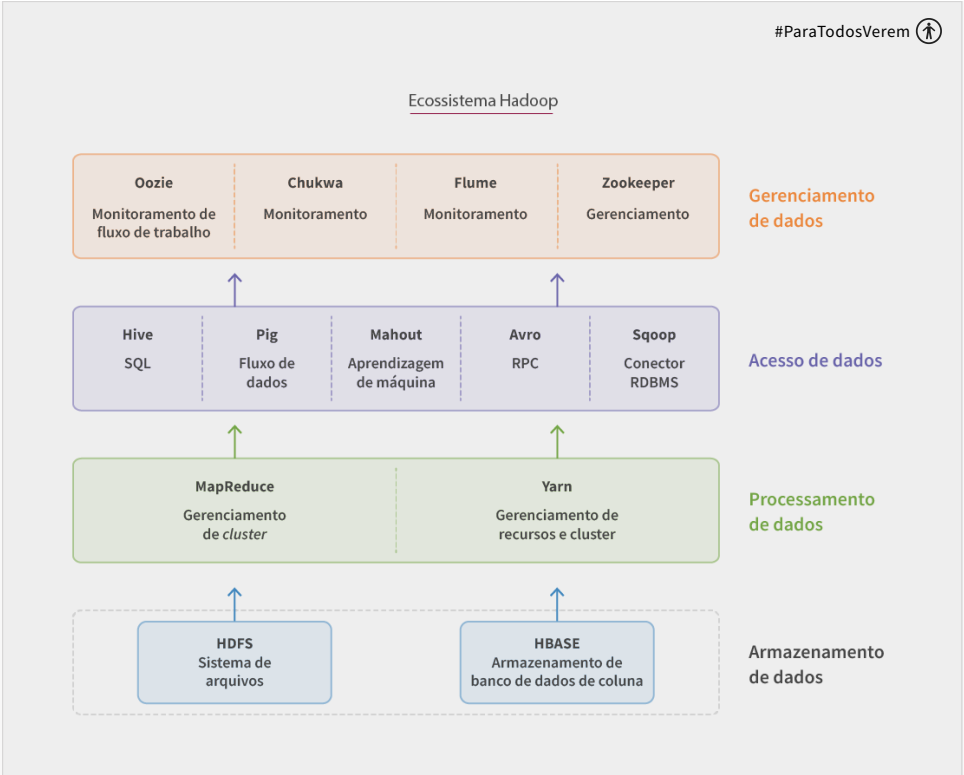


Figura 3: Ecossistema Hadoop. Fonte: Adaptado de Dekate, 2023.

**Código aberto: Liberdade e flexibilidade:**

O Hadoop é um projeto da **Apache Software Foundation (ASF)**, o que significa que seu código é **gratuito e aberto**. Você pode:

- Acessar o código-fonte e **customizá-lo** de acordo com suas necessidades.
- **Contribuir para o desenvolvimento** do projeto, tornando-o ainda mais poderoso.

**Comunidade vibrante: apoio e colaboração:**

Uma das maiores vantagens do Hadoop é sua **comunidade ativa** de desenvolvedores e usuários. Isso significa que você encontrará:

- **Documentação extensa** e tutoriais para aprender sobre o Hadoop.
- **Fóruns *on-line*** e grupos de discussão para tirar dúvidas e compartilhar conhecimentos.
- **Suporte profissional** de empresas que oferecem serviços e soluções Hadoop.

**Aplicações do Hadoop:**

O Hadoop é usado em diversos setores para:

- **Análise de *logs*** para identificar padrões e anomalias.
- **Análise de sentimento** para entender a opinião pública sobre produtos, serviços ou eventos.
- **Recomendação de produtos** para oferecer aos clientes experiências personalizadas.
- **Detecção de fraudes** para proteger sistemas contra atividades maliciosas.
- **Análise de séries temporais** para prever tendências e tomar decisões mais inteligentes.

## Apache Spark: Uma Visão Geral

O Spark é um *framework* de processamento de alto desempenho para grandes conjuntos de dados. Oferece um modelo de programação unificado e é adequado para várias tarefas, como processamento em lote, *streaming* de dados e *machine learning*. O Spark é composto de diversos componentes que trabalham em conjunto para oferecer alto desempenho e escalabilidade. Na figura 4 apresentamos a arquitetura do Spark e os respectivos componentes que a compõe:

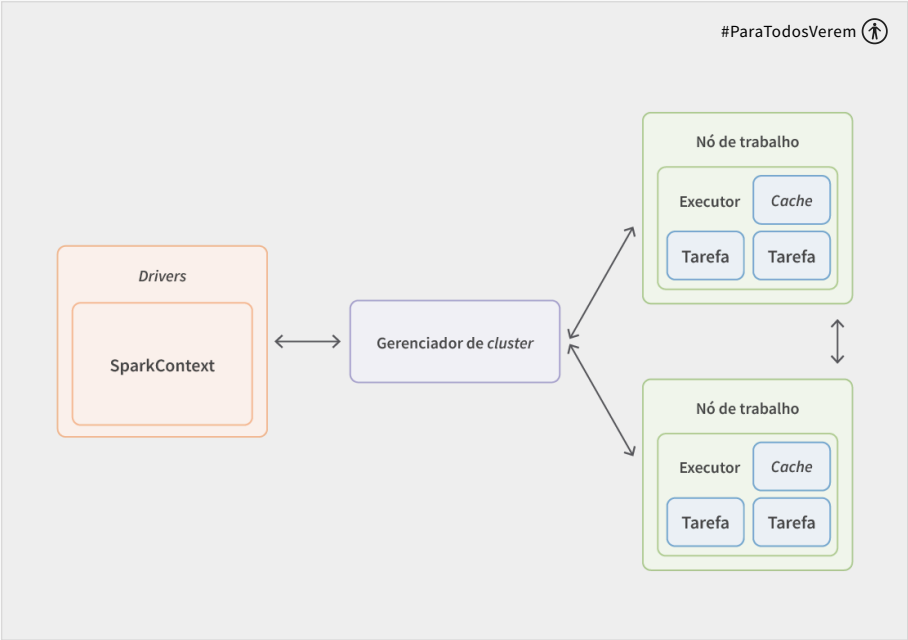


Figura 4: Arquitetura de *cluster* no Spark. Fonte: Adaptado de Pereira, 2020.

Nesta arquitetura temos:

- ***Driver*:** responsável por gerenciar o *cluster* e coordenar as tarefas.
- ***Workers*:** responsáveis por executar as tarefas no cluster.
- **RDDs (Resilient Distributed Datasets):** Coleções de dados particionados e distribuídos na memória do cluster.

O Spark foi desenvolvido visando a abranger uma ampla gama de tarefas, desde processamento em lotes até *streaming*, tornando mais simples e econômico combinar diferentes tipos de processamento em um único mecanismo. Isso é fundamental para estruturas de análise de dados em produção, evitando a necessidade de gerenciar e manter várias ferramentas separadas. Além disso, ele se integra facilmente com outras ferramentas de *big data* conforme necessário. E como o Spark faz isso? Por meio de vários componentes que servem como um motor computacional fornecendo funções básicas como mapeamento, redução, filtro e coleta de dados. Essas funcionalidades são essenciais para um processamento de dados eficiente.

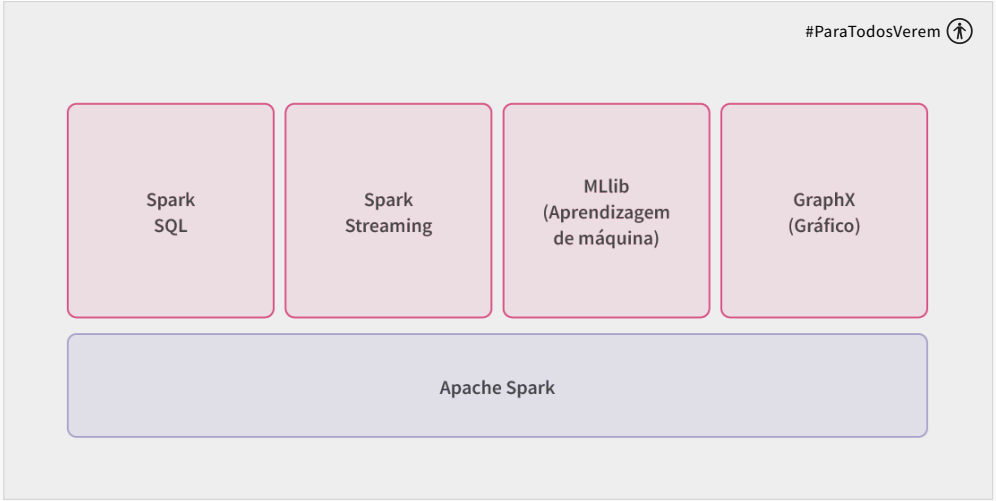


Figura 5: Ferramentas de alto nível no Spark. Fonte: Adaptado de Pereira, 2020.

Em relação às demais ferramentas que compõem o Spark, temos:

- **Spark Core:** módulo central que fornece as funcionalidades básicas do Spark.
- **Spark SQL:** módulo para SQL e análise de dados estruturados.
- **Spark Streaming:** módulo para processamento de dados em tempo real.
- **Spark MLlib:** módulo para Machine Learning.

**Quais as vantagens do Spark?**

O Spark é um *framework* de processamento de *big data* de código aberto que oferece diversas vantagens para desenvolvedores e cientistas de dados. Entre as principais vantagens, podemos destacar:

- **Alto desempenho:** o Spark é otimizado para processamento em memória, o que oferece alto desempenho para diversas tarefas.
- **Escalabilidade:** o Spark pode ser facilmente escalado para lidar com grandes conjuntos de dados em clusters de commodity.
- **Modelo de programação unificado:** o Spark oferece um único modelo de programação que pode ser usado para diversos tipos de tarefas.
- **Rica biblioteca de APIs:** o Spark oferece uma ampla biblioteca de APIs para diversas tarefas de Big Data.
- **Comunidade ativa:** o Spark possui uma comunidade ativa de desenvolvedores e usuários que contribuem para o desenvolvimento da ferramenta.

**Quando usar o Spark?**

O Spark é uma ferramenta ideal para diversos tipos de tarefas de *big data*, como:

- **Análise de grandes conjuntos de dados:** processamento em lote, *streaming* de dados etc.
- **Machine Learning:** treinamento de modelos, inferência etc.
- **Aplicações em tempo real:** análise de eventos, fraudes etc.

## Comparação entre Hadoop e Spark

No quadro 2, apresentamos uma comparação entres características do Hadoop em relação às do Spark:

| Característica | Hadoop | Spark |
|---|---|---|

| | | |
|---|---|---|
| Modelo de programação | MapReduce | SQL, *streaming*, *machine learning* |
| Processamento | Disco | Memória |
| Desempenho | Baixo a médio | Alto |
| Complexidade | Alta | Média |

Enquanto o Hadoop utiliza o modelo de programação MapReduce e processa dados principalmente a partir do disco, o Spark adota uma abordagem mais versátil, suportando SQL, *streaming* e aprendizado de máquina, realizando o processamento em memória. Isso confere ao Spark um desempenho superior ao Hadoop. Além disso, a complexidade de utilização tende a ser menor no Spark em comparação com o Hadoop, tornando-o mais acessível e amigável para uma variedade de aplicações e usuários.

## Quando Utilizar Cada Ferramenta

Hadoop e Spark são *frameworks* de *big data* amplamente utilizados que oferecem diferentes funcionalidades e são adequados para diferentes tipos de tarefas. Aqui está um guia para ajudá-lo a escolher a ferramenta certa para o seu projeto:

### Hadoop

- Ideal para processamento em lote de grandes volumes de dados.
- Suporta tarefas como ETL (*Extract, Transform, Load*), análise de dados e geração de relatórios.
- Oferece alta escalabilidade e confiabilidade.
- É uma boa escolha para projetos que exigem processamento de grandes conjuntos de dados históricos.

### Spark

- Ideal para processamento em tempo real, *streaming* e *machine learning*.
- Suporta tarefas como análise de *streaming* de dados, *machine learning* em larga escala e análise de grafos.
- Oferece alto desempenho e facilidade de uso.
- É uma boa escolha para projetos que exigem processamento de dados em tempo real ou *machine learning* em larga escala.

Voltando ao exemplo da Netflix que mencionamos na introdução, podemos ver claramente como o uso eficiente do *big data*, por meio de ferramentas como o Apache Spark e o Hadoop, revolucionou a maneira como consumimos entretenimento. A capacidade de entender as preferências individuais dos usuários e oferecer recomendações personalizadas não apenas melhora a experiência do usuário, mas também impulsiona os negócios da empresa.

Além disso, o caso da Netflix nos leva a refletir sobre o potencial das ferramentas de *big data* para resolver problemas complexos em diversas áreas, como saúde, finanças, transporte, entre outros. Imagine o impacto de utilizar dados em larga escala para prever surtos de doenças, detectar fraudes financeiras em tempo real ou otimizar o tráfego urbano.

Essas reflexões nos levam a considerar: até onde podemos chegar utilizando as ferramentas de *big data*? Como podemos aproveitar ao máximo essas tecnologias para enfrentar os desafios do mundo moderno? E, acima de tudo, como garantir que essas ferramentas sejam utilizadas de forma ética e responsável?

Essas são questões importantes que devemos continuar explorando à medida que avançamos no universo do Big Data. Afinal, o potencial é imenso, mas é fundamental que usemos essas poderosas ferramentas com sabedoria e em benefício da sociedade como um todo.

## Conclusão

À medida que encerramos nossa jornada pela visão geral dos *frameworks* de *big data*, é evidente que adquirimos um entendimento mais profundo sobre a importância dessas ferramentas na era da informação. Desde a compreensão dos desafios do *big data* até a exploração das funcionalidades dos principais de *frameworks* como Hadoop e Spark, cada passo nos aproximou mais do vasto potencial que os dados oferecem. Recordemos que, embora tenhamos percorrido apenas uma parte desse universo em constante expansão, estamos agora equipados com os conhecimentos essenciais para navegar nesse terreno desafiador. Compreendemos os desafios, apreciamos as oportunidades e estamos prontos para explorar novos horizontes. Que essa visão geral seja o ponto de partida para uma jornada emocionante e frutífera no mundo dos dados. Obrigado por embarcar conosco nessa exploração – o futuro está nas nossas mãos, e os dados são o nosso guia. Até a próxima aventura!

## Referências Bibliográficas


DEKATE, C. Introduction to Hadoop architecture and its components. **Analytics Vidhya**. Disponível em: https://www.analyticsvidhya.com/blog/2022/06/introduction-to-hadoop-architecture-and-its-components/. Acesso em: 20 mar. 2024.

LIN, H.; HARDING, J; CHEN, C. **A hyperconnected manufacturing collaboration system using the semantic web and hadoop ecosystem system**. [*S.l.*]: Procedia CIRP, 2016.

MARQUESONE, R. **Big data**: técnicas e tecnologias para extração de valor dos dados. 2. ed. São Paulo: Editora Casa do Código, 2016.

NETFLIX TECH BLOG. **Hadoop**. Disponível em: https://netflixtechblog.com/tagged/hadoop. Acesso em: 20 mar. 2024.

PEREIRA, M. A.; NEUMANN, F. B.; MILANI, A. M. P. *et al*. **Framework de big data**. Porto Alegre: Grupo A, 2020.

SANTOS, J. O. Hadoop: seus componentes principais e sua evolução. **Medium**, 2023. Disponível em: https://johnosd.medium.com/hadoop-seus-componentes-principais-e-sua-evolu%C3%A7%C3%A3o-cf125c99fadd. Acesso em: 20 mar. 2024.