

Big Data



Fundamentos 3.0

Data Science Academy



Data Science Academy

Data Science Academy

A Data Science Academy é um portal de ensino online especializado em Big Data, Machine Learning, Inteligência Artificial, Blockchain, RPA e tecnologias relacionadas.

Nosso objetivo é fornecer aos alunos conteúdo de alto nível por meio do uso de computador, tablet ou smartphone, em qualquer lugar, a qualquer hora, 100% online e 100% em português.

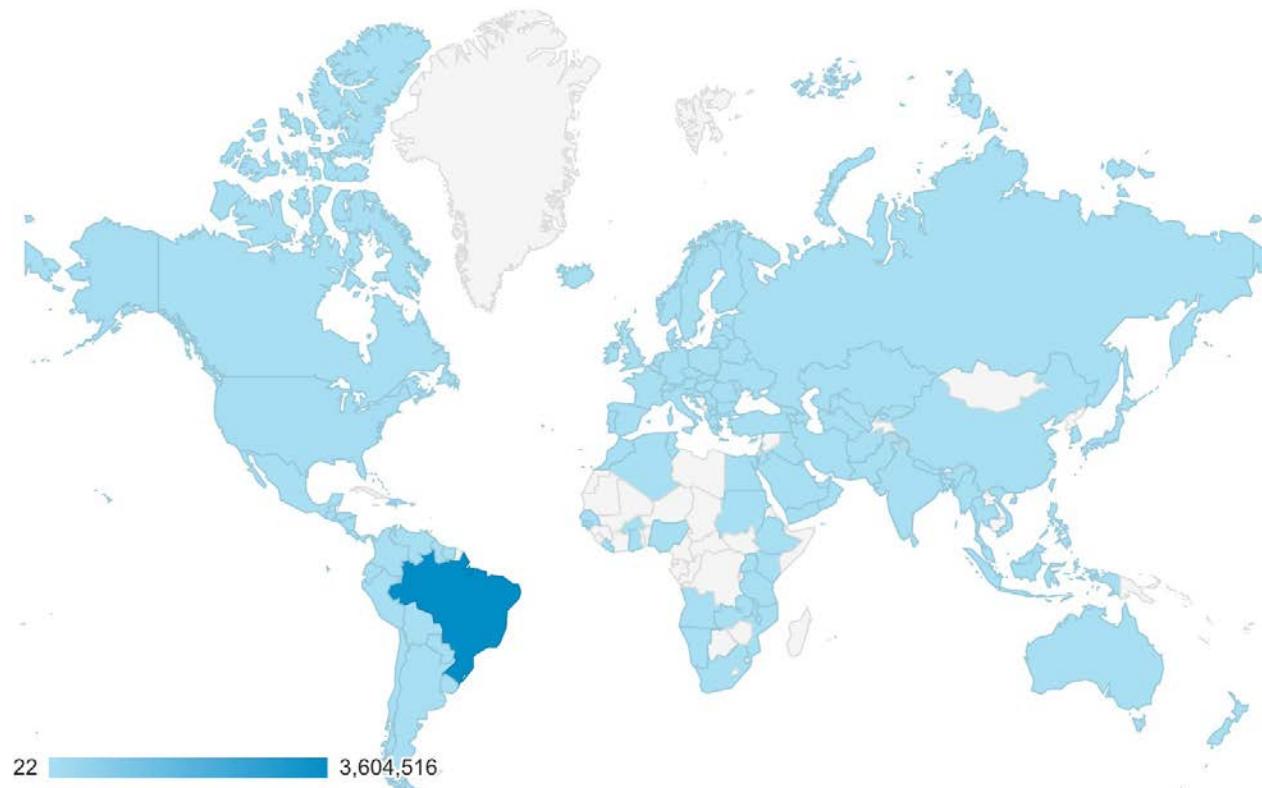
Nossa
Escola





Data Science Academy - Localização

No Brasil e no Mundo.





Conteúdo Programático



Capítulo 1

Introdução



Capítulo 2

O Que é Big Data?



Capítulo 3

Sistemas de Armazenamento de Dados



Capítulo 4

Armazenamento e Processamento Paralelo



Capítulo 5

Cloud Computing



Capítulo 6

MLOps e DataOps



Capítulo 7

Dados Como Serviço



Capítulo 8

ETL - Extração, Transformação e Carga de Dados



Capítulo 9

Como Iniciar Um Projeto de Big Data



Capítulo 10

Avaliação Final e Certificado de Conclusão

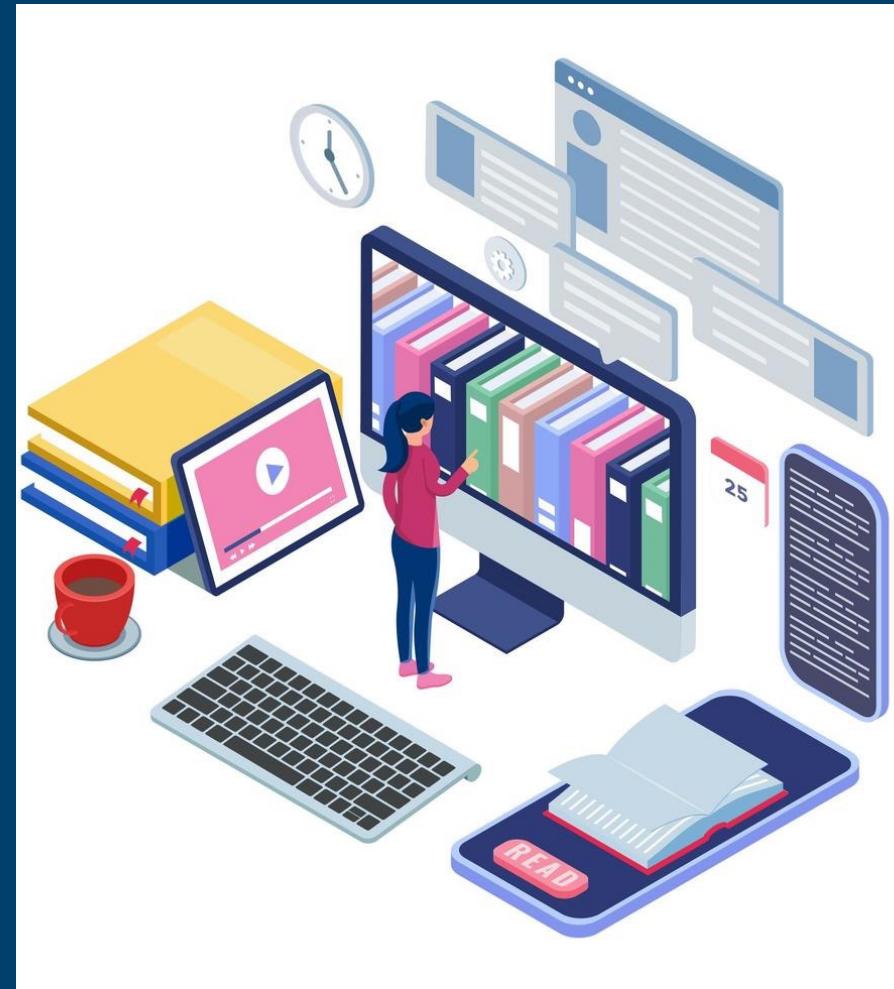


Big Data Fundamentos 3.0

**Avaliação Final,
Certificado de
Conclusão e E-book do
Curso**

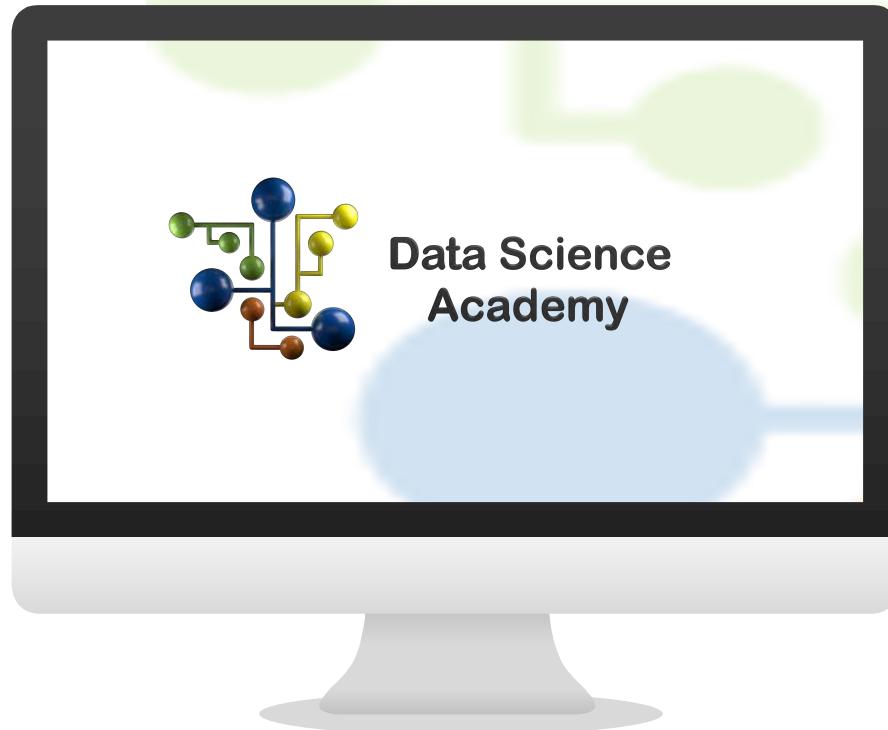
Data Science Academy

Data Science Academy

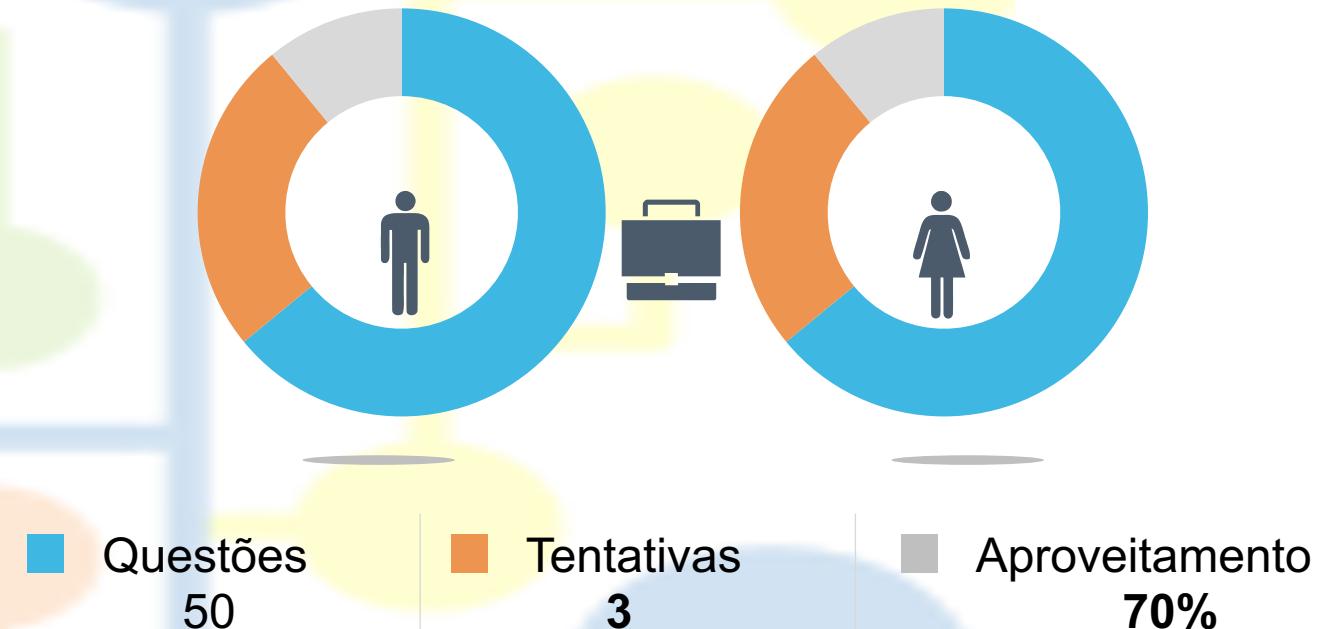




Avaliação Final, Certificado de Conclusão e E-book do Curso



Data Science
Academy

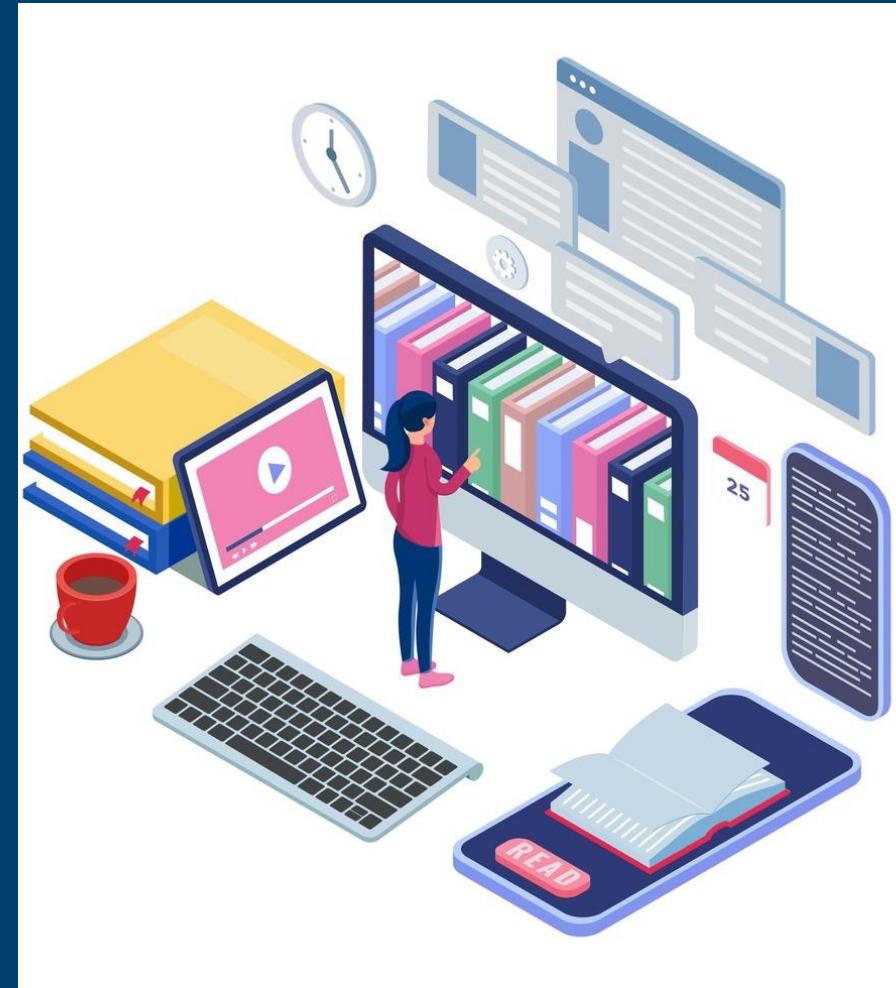


Data Science Academy

Big Data Fundamentos 3.0

Qual o PÚblico Alvo Deste Curso?

Data Science Academy





Qual o PÚblico Alvo Deste Curso?

Este é um curso teórico. Apenas relaxe e aproveite o conhecimento que será trazido a você.

Qualquer pessoa interessada em aprender sobre o universo do Big Data pode acompanhar este curso!

Este curso é pré-requisito recomendado para todos os demais cursos em nosso portal.



Big Data

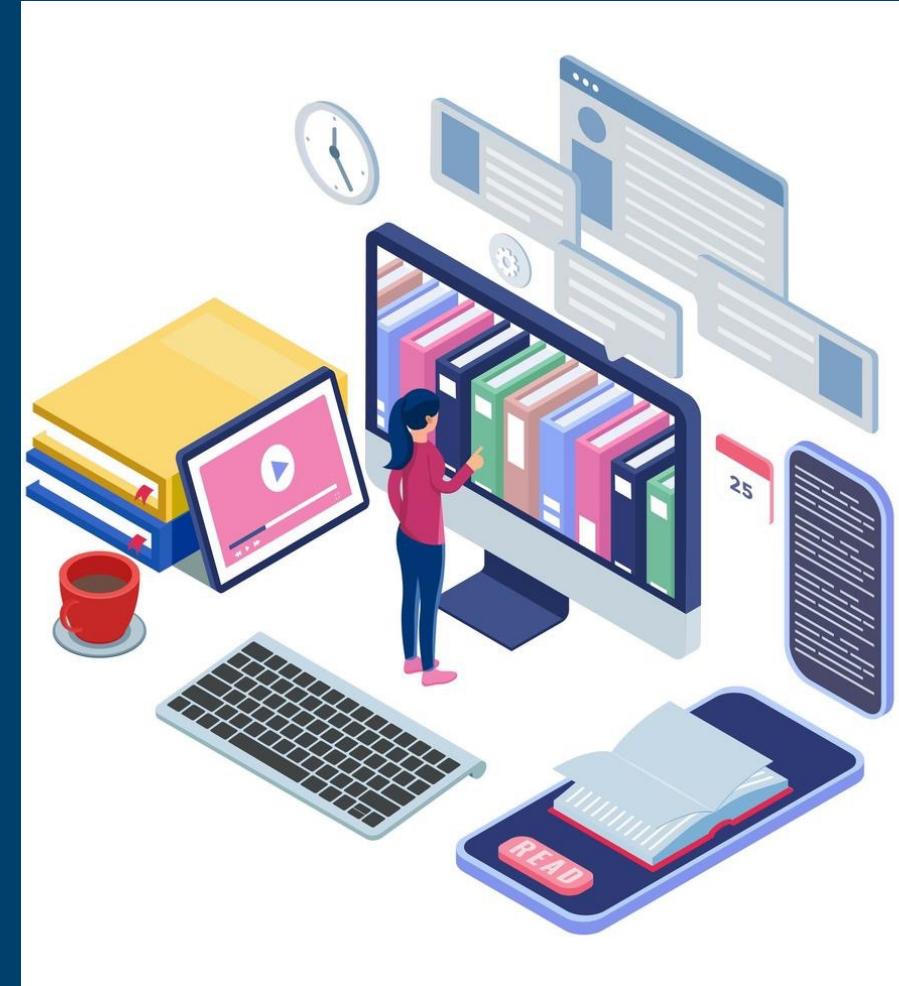
Fundamentos 3.0

Introdução

O Que é Big Data?

Data Science Academy

Data Science Academy



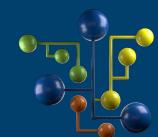
Big Data

Fundamentos 3.0

Fatos Sobre o Big Data

Data Science Academy

Data Science Academy





Big Data

Neste exato momento, uma verdadeira enxurrada de dados, ou 2.5 quintilhões de bytes por dia, é gerada para nortear indivíduos, empresas e governos, e está dobrando a cada dois anos.





Fatos Sobre o Big Data

Big Data

Cerca de 90% de todos os dados gerados no planeta, foram gerados nos últimos 2 anos.





Big Data

Aproximadamente 80% dos dados são não-estruturados ou estão em diferentes formatos, o que dificulta a análise.





Fatos Sobre o Big Data

Big Data

Toda vez que fazemos uma compra, uma ligação ou interagimos nas redes sociais, estamos produzindo esses dados.



Fatos Sobre o Big Data

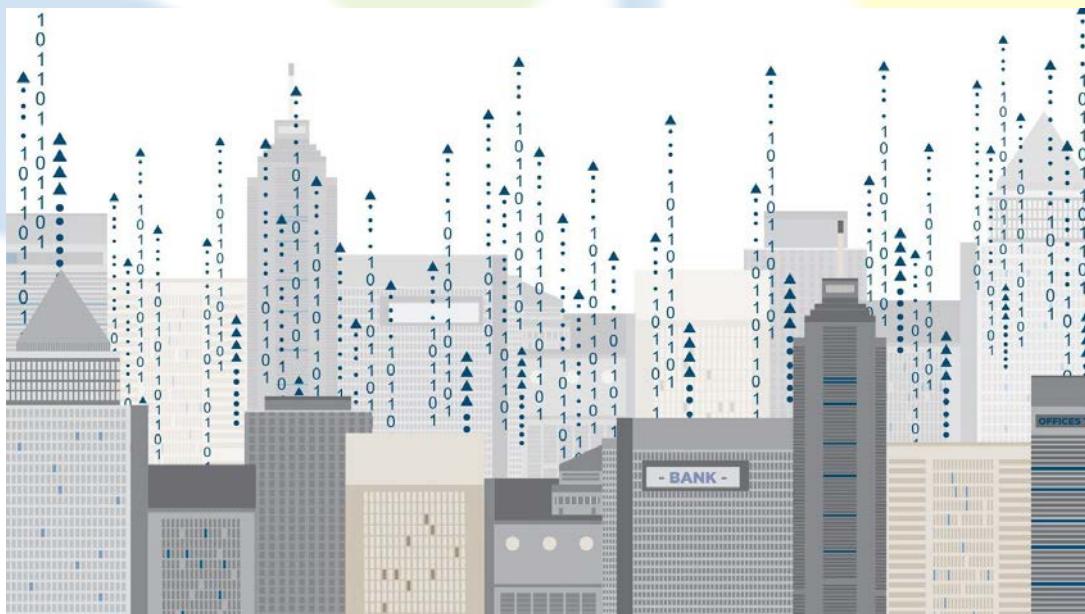
E com a recente conectividade em objetos, tal como relógios, carros e até geladeiras, as informações capturadas se tornam massivas e podem ser cruzadas para criar modelos preditivos cada vez mais elaborados, apontando e, até prevendo, o comportamento de empresas e clientes.





O Que é Big Data?

Big Data é uma coleção de conjuntos de dados, grandes e complexos, que não podem ser processados por bancos de dados ou aplicações de processamento tradicionais.



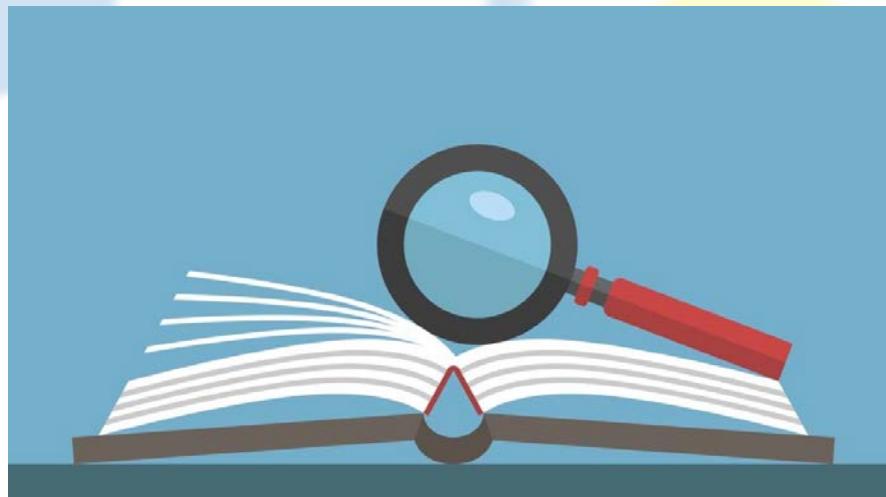
O Que é Big Data?

O Google estima que a humanidade criou nos últimos 5 anos, o equivalente a 300 Exabytes de dados ou seja:
300.000.000.000.000.000 bytes de dados.



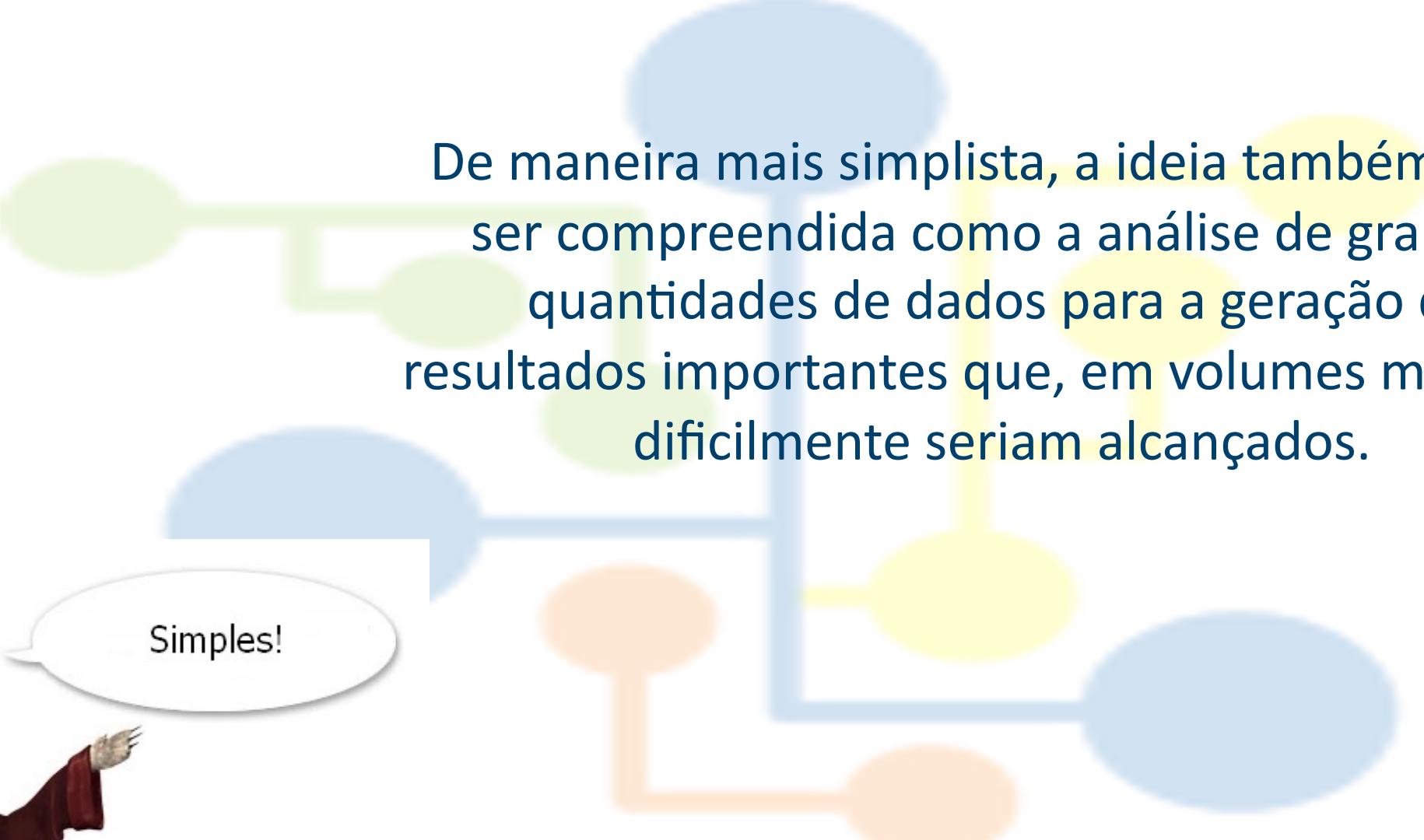
O Que é Big Data?

Podemos definir o conceito de Big Data como sendo conjuntos de dados extremamente amplos e que, por este motivo, necessitam de ferramentas especialmente preparadas para lidar com grandes volumes, velocidade e variedade, de forma que toda e qualquer informação disponível nos dados possa ser encontrada, analisada e aproveitada em tempo hábil.





O Que é Big Data?



De maneira mais simplista, a ideia também pode ser compreendida como a análise de grandes quantidades de dados para a geração de resultados importantes que, em volumes menores, dificilmente seriam alcançados.



O Que é Big Data?

O Big Data nos dá uma visão clara do que é granular!



O Que é Big Data?

No mundo do Big Data não temos de nos fixar na causalidade; podemos descobrir padrões e correlações nos dados que nos propiciem novas e valiosas ideias.



Big Data

Fundamentos 3.0

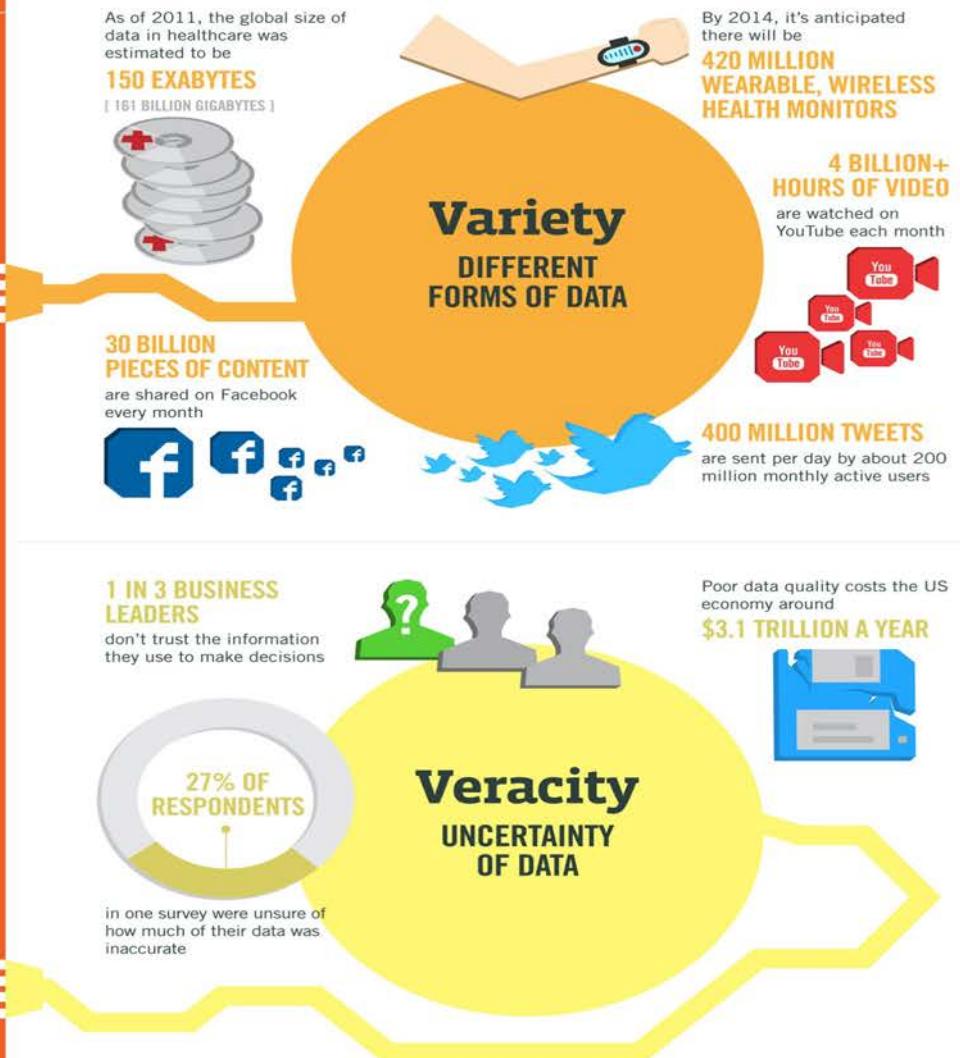
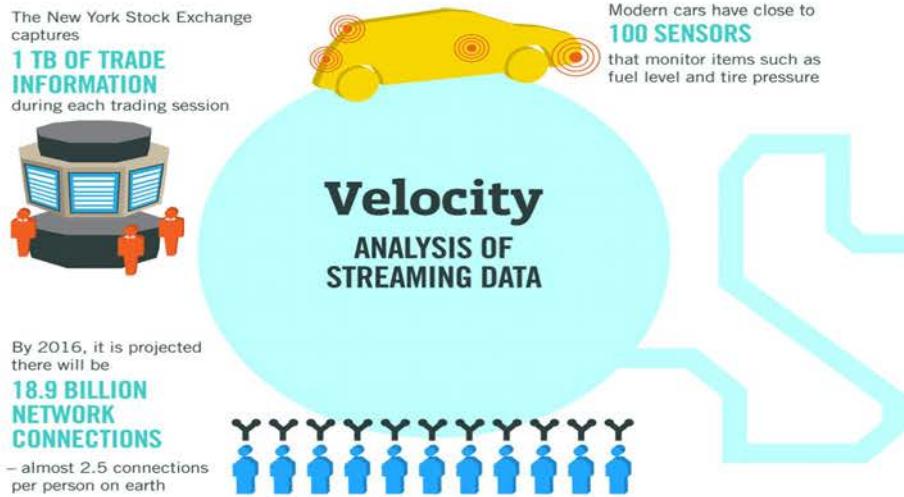
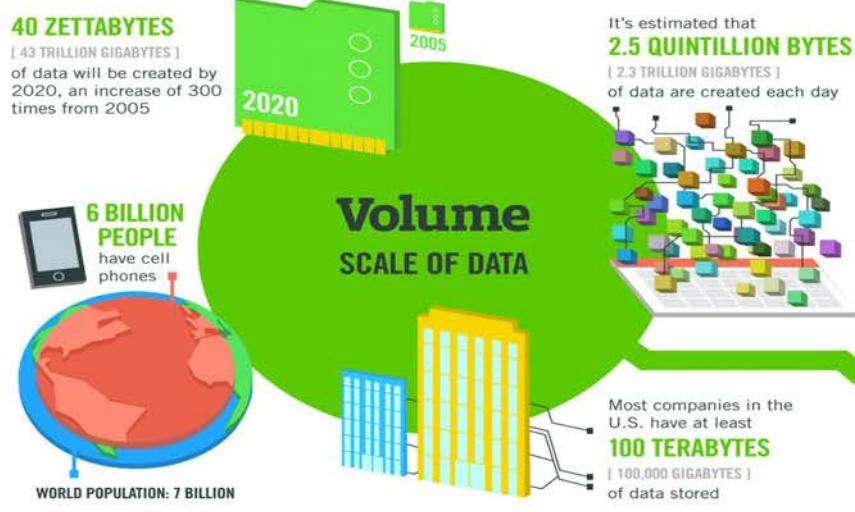
Os 4 V's do Big Data

Data Science Academy

Data Science Academy



Os 4Vs do Big Data



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM



Os 4 V's do Big Data

Volume

Tamanho dos Dados.

Velocidade

Geração dos Dados.

Variedade

Formato dos Dados

Veracidade

Confiabilidade dos Dados





Os 4 V's do Big Data

Volume

Tamanho dos Dados.

- Espera-se que 40 zettabytes de dados sejam criados por ano no mundo;
- Cerca de 2.5 quintillionbytes de dados são criados por dia;
- Existem atualmente cerca de 6 bilhões de telefones móveis no planeta;
- Cada empresa americana armazena cerca de 100 Terabytes de dados.





Os 4 V's do Big Data

Variedade

Formato dos Dados.

- 150 exabytes é a estimativa de dados que foram gerados especificamente para tratamento de casos de doença em todo o mundo por ano desde 2011;
- Mais de 4 bilhões de horas por mês são usadas para assistir vídeos no YouTube;
- 30 bilhões de imagens são publicadas por mês no Facebook;
- 200 milhões de usuários ativos por mês, publicam 400 milhões de tweets por dia.





Os 4 V's do Big Data

Velocidade

Geração dos Dados.

- 1 terabyte de informação é criada durante uma única sessão da bolsa de valores Americana, a New York Stock Exchange (NYSE);
- Aproximadamente 100 sensores estão instalados nos carros modernos para monitorar nível de combustível, pressão dos pneus e muitos outros aspectos do veículo;
- 18.9 bilhões de conexões de rede já existem no mundo.





Os 4 V's do Big Data

Veracidade

Confiabilidade dos Dados.

→ Atualmente, 1 em cada 3 gestores tem experimentado problemas relacionados a veracidade dos dados para tomar decisões de negócios.

→ Além disso, estima-se que 3.1 trilhões de dólares por ano sejam desperdiçados devido a problemas de qualidade dos dados.



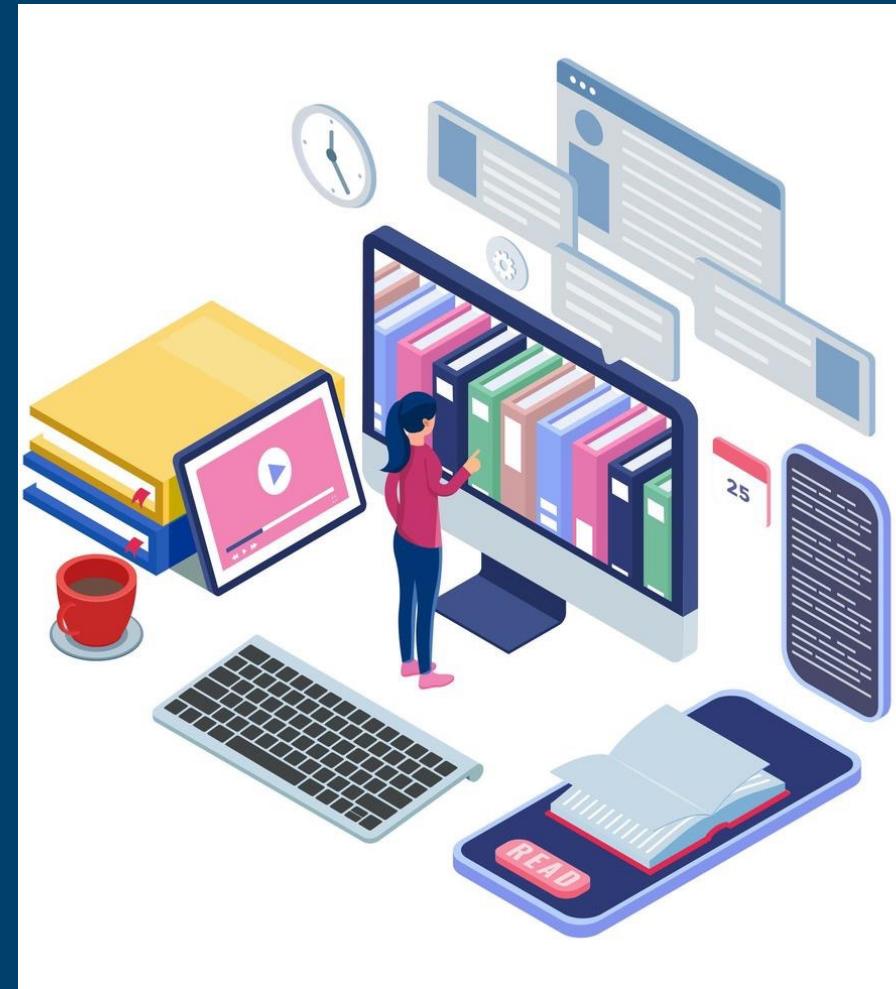
Big Data

Fundamentos 3.0

Big Data x Ciência de Dados

Data Science Academy

Data Science Academy





Big Data x Ciência de Dados

Big Data e Ciência de Dados são a mesma coisa?

Não

Big Data é a matéria-prima, ou seja, dados.
Ciência de Dados é um conjunto de técnicas para análise de dados.

Quando aplicamos Ciência de Dados ao Big Data extraímos valor e
então temos o que é chamado de **Big Data Analytics**.



Big Data Fundamentos 3.0

Exemplos de Aplicação do Big Data Analytics

Data Science Academy

Data Science Academy





Exemplos de Aplicação do Big Data Analytics

Uma Companhia Área pode extrair, armazenar, processar e analisar dados de viagens dos passageiros a fim de oferecer rotas com maior probabilidade de venda.

Uma Rede de Supermercados pode extrair, armazenar, processar e analisar dados de compras a fim de detectar padrões e organizar os produtos de forma a aumentar as vendas.

Uma Rede de Hotéis pode extrair, armazenar, processar e analisar dados de comentários de clientes em redes sociais para customizar seus serviços, aumentar as vendas e reduzir custos.

Uma Rede de Hospitais pode extrair, armazenar, processar e analisar dados de exames médicos a fim de personalizar e otimizar o atendimento dos pacientes.



Big Data

Fundamentos 3.0

Sistemas de Armazenamento de Dados

Data Science Academy

Data Science Academy



Big Data

Fundamentos 3.0

O V de Volume em Big Data

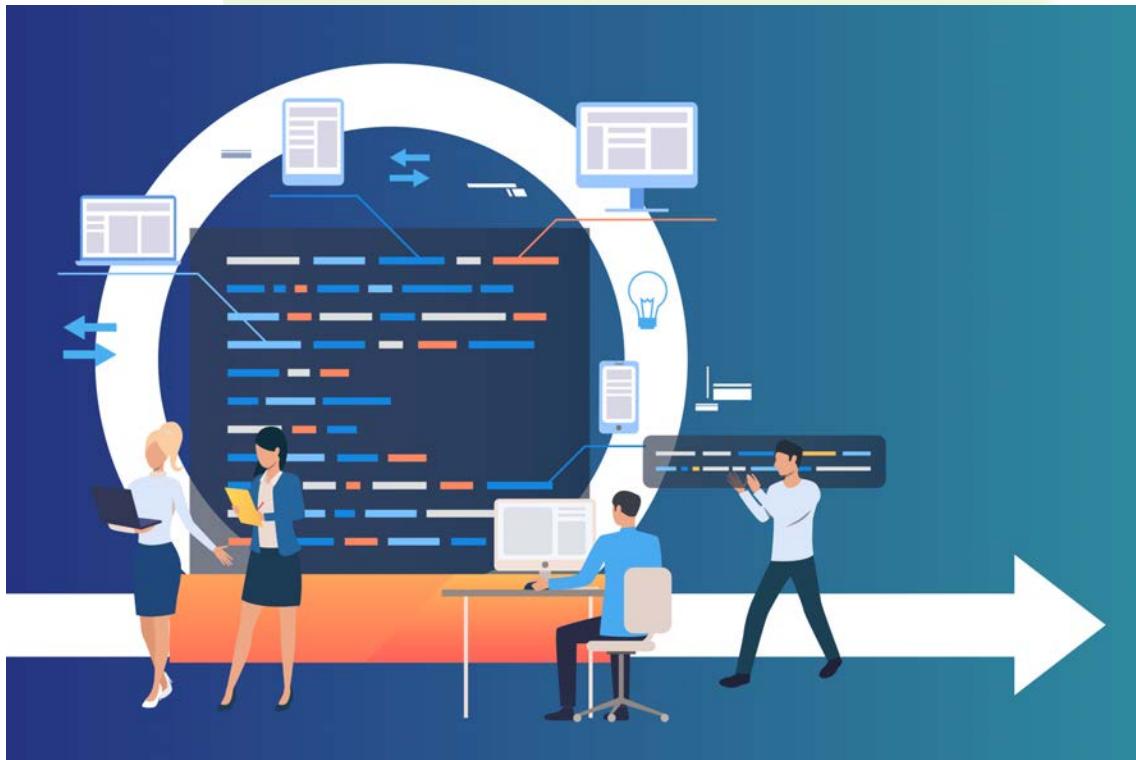
Data Science Academy

Data Science Academy





O V de Volume em Big Data



O V de Volume é crítico em Big Data.

Como vamos armazenar grandes conjuntos de dados?

Como vamos acessar grandes conjuntos de dados armazenados?

Precisamos realmente armazenar tudo?



Big Data Fundamentos 3.0

Como Armazenamos Big Data?

Data Science Academy

Data Science Academy





Como Armazenamos Big Data?

Em linhas gerais o armazenamento pode ser feito com base na seguinte regra:

Os dados são estruturados ou podem ser estruturados antes do armazenamento?

Usamos um Data Warehouse!

Os dados **NÃO** são estruturados ou **NÃO** podem ser estruturados antes do armazenamento?

Usamos um Data Lake ou um Data Store!



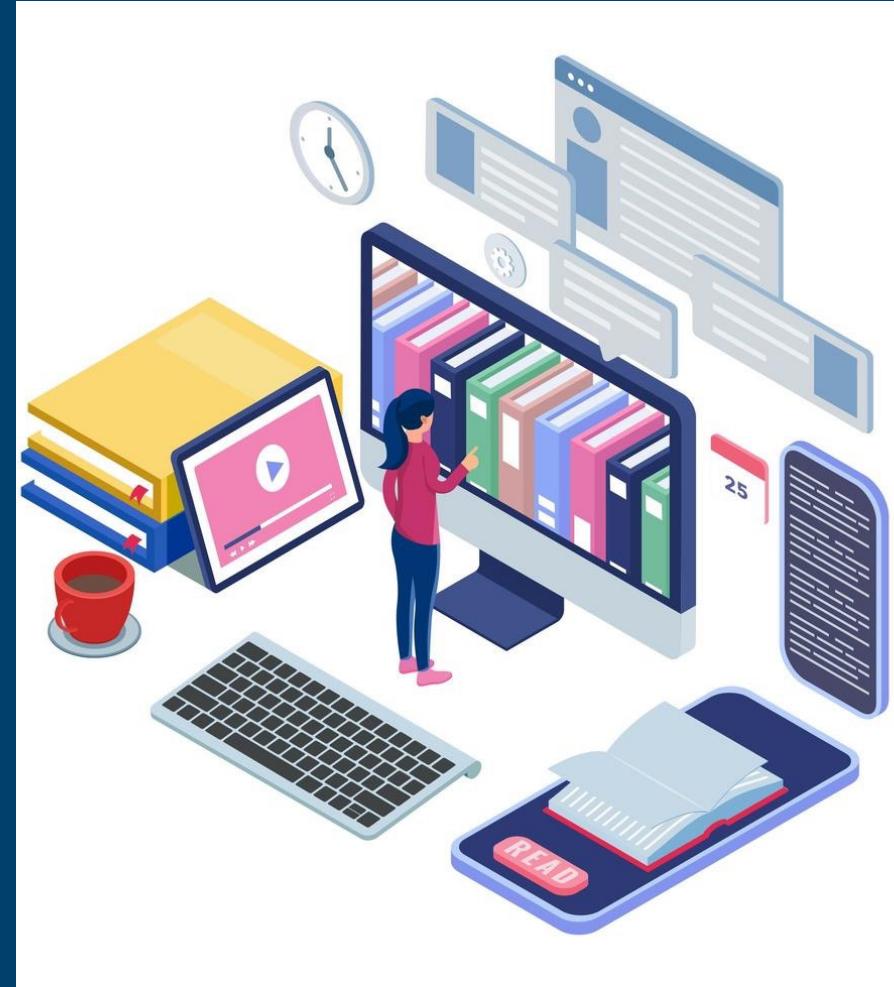
Big Data

Fundamentos 3.0

Bancos de Dados Relacionais x Bancos de Dados NoSQL

Data Science Academy

Data Science Academy



Bancos de Dados Relacionais x Bancos de Dados NoSQL

Bancos de Dados Relacionais são bancos de dados estruturados e com schema (organização dos dados) bem definido.

O schema é definido e criado antes do armazenamento dos dados.

Um Data Warehouse, por exemplo, é criado com alguma tecnologia de banco relacional como SGBD (Sistema Gerenciador de Banco de Dados) Oracle, IBM DB2, Microsoft SQL Server, MySQL, PostgreSQL e muitos outros.

Em um banco de dados relacional os dados são organizados em tabelas que se relacionam.



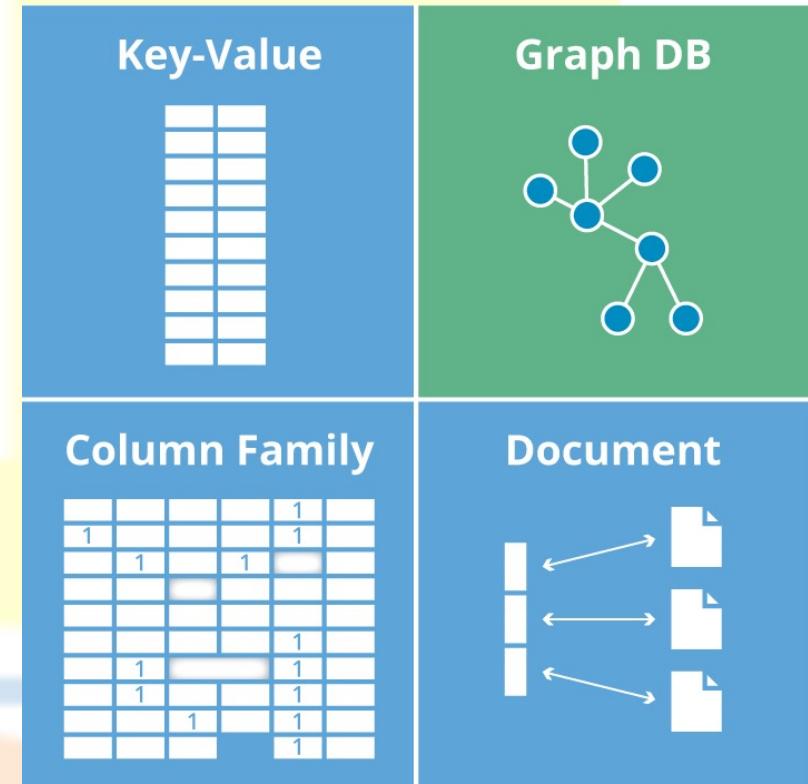
Bancos de Dados Relacionais x Bancos de Dados NoSQL

Bancos de Dados Não Relacionais (NoSQL) partem do princípio que os dados podem ser semi ou não estruturados e que outros tipos de relacionamentos podem existir entre os dados.

Podemos usar Bancos de Dados Não Relacionais (NoSQL) para construir Data Lakes e Data Stores (Data Lakes e Data Stores são conceitos, como veremos mais adiante).

Normalmente não precisamos definir o schema antes do armazenamento ou o schema é definido no momento do armazenamento dos dados.

Existem diversos tipos de bancos de dados NoSQL.



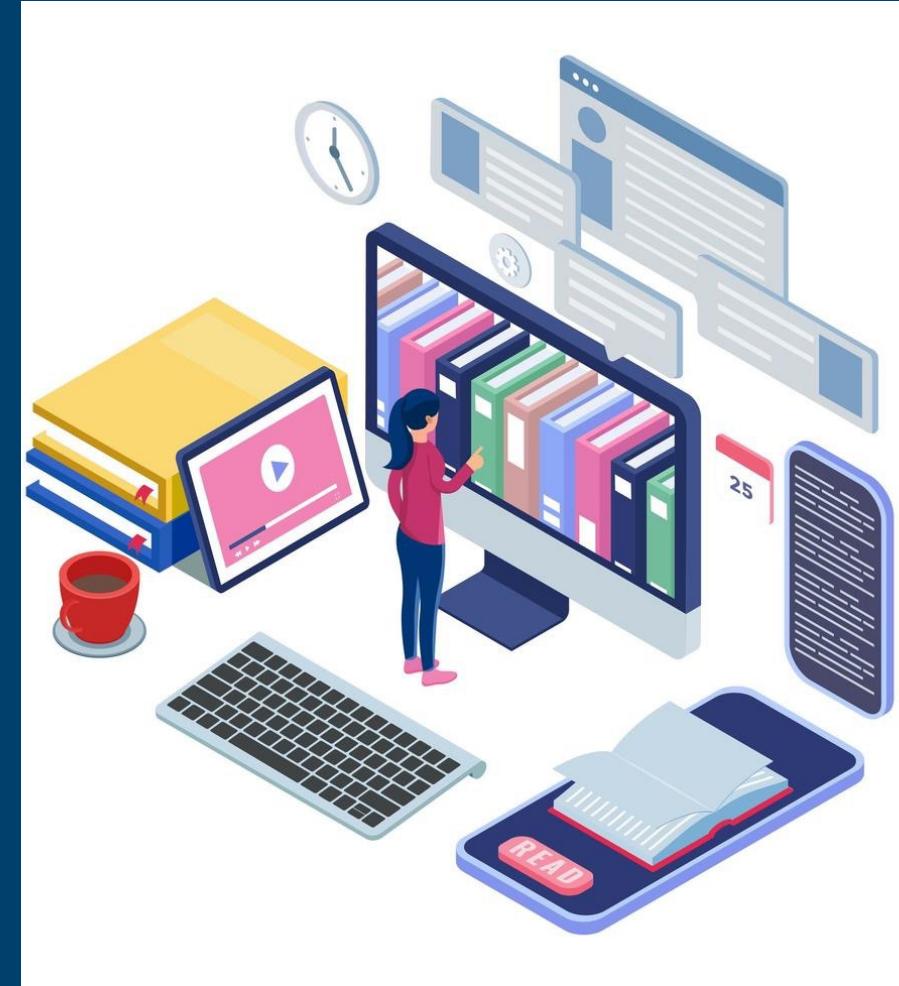
Big Data

Fundamentos 3.0

Definindo Data Warehouses

Data Science Academy

Data Science Academy



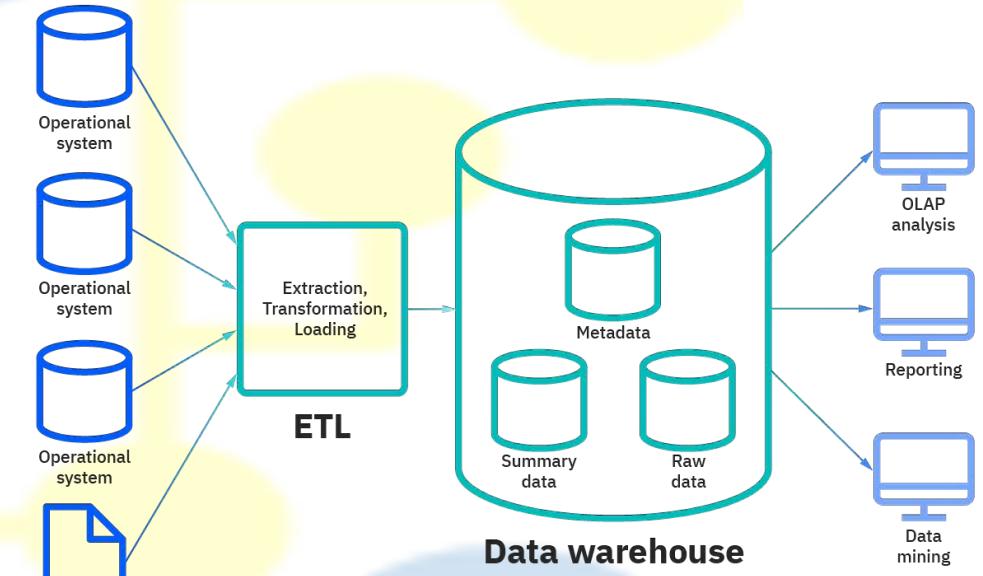


Definindo Data Warehouses

Um Data Warehouse (DW) é um sistema de armazenamento que conecta e harmoniza grandes quantidades de dados de muitas fontes diferentes.

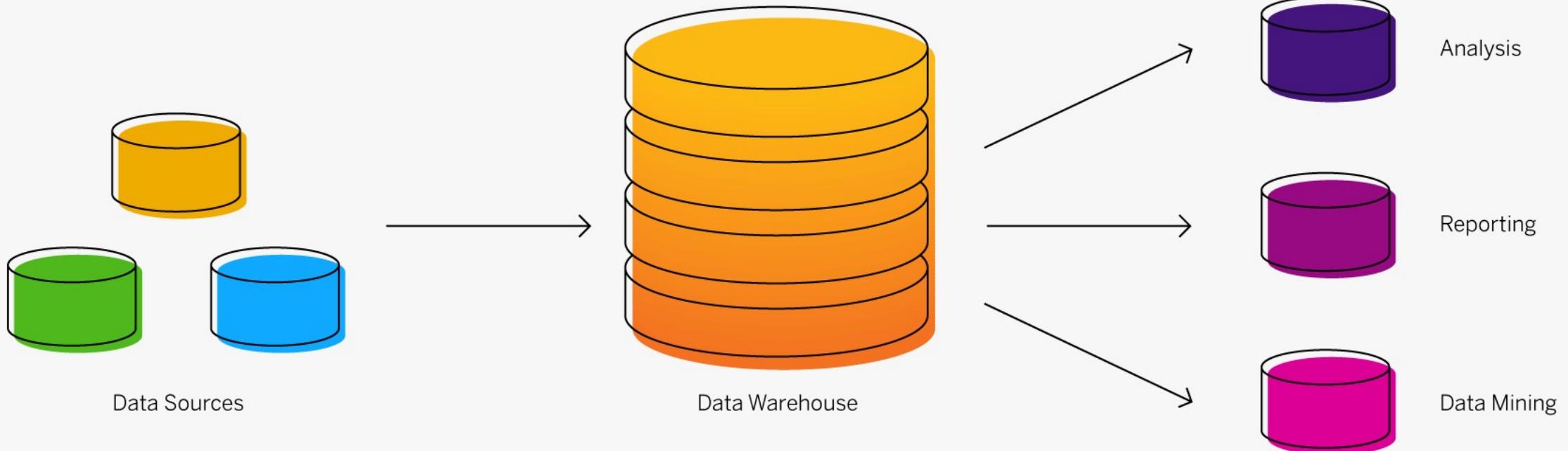
Seu objetivo é alimentar a inteligência de negócios (Business Intelligence), relatórios e análises e oferecer suporte aos requisitos de negócio, para que as empresas possam transformar seus dados em insights e tomar decisões inteligentes baseadas em dados.

Os DWs armazenam dados atuais e históricos em um único lugar e atuam como a única fonte de informações confiáveis para uma organização.





Definindo Data Warehouses



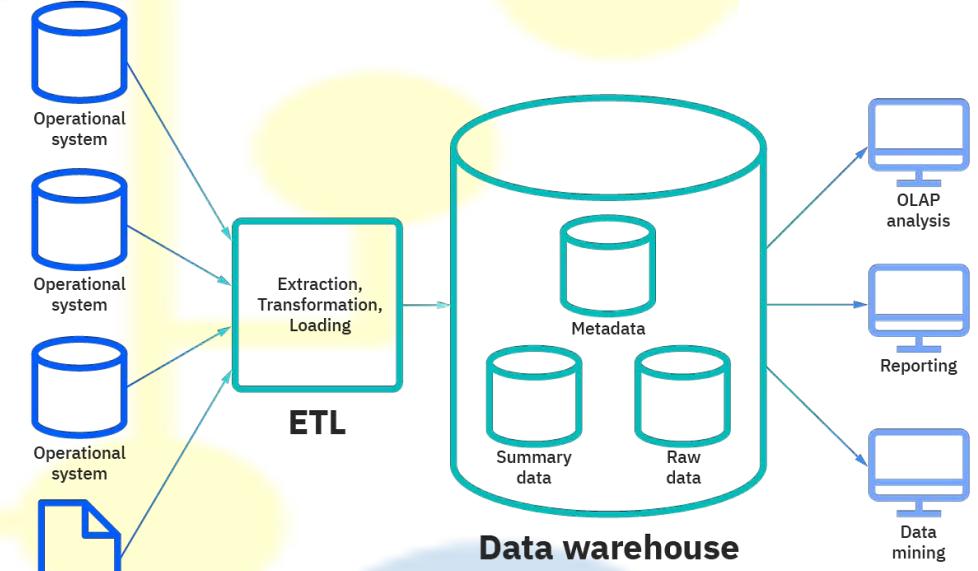
Definindo Data Warehouses

Os dados fluem para um DW a partir de sistemas transacionais (como ERP e CRM), bancos de dados e fontes externas, como sistemas de parceiros, dispositivos de Internet das Coisas (IoT), aplicativos de mídia social - geralmente em uma cadência regular.

O surgimento da computação em nuvem causou uma mudança no cenário.

Nos últimos anos, os locais de armazenamento de dados mudaram da infraestrutura local tradicional para vários locais, incluindo nuvem privada e nuvem pública.

O schema deve ser definido antes do processo de armazenamento dos dados.



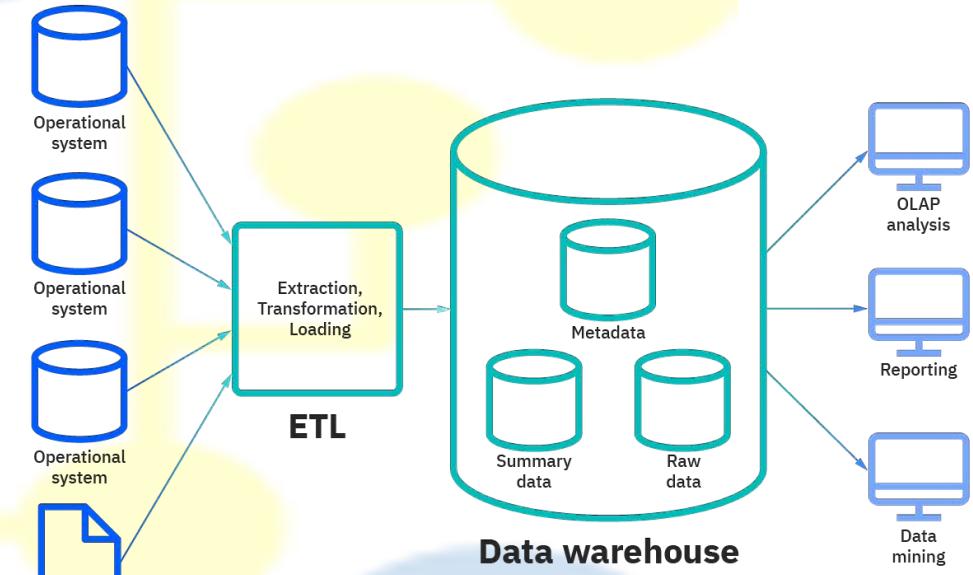


Definindo Data Warehouses

Os DWs modernos são projetados para lidar com dados estruturados e não estruturados, como vídeos, arquivos de imagem e dados de sensor (embora Data Lakes ainda sejam opções melhores para dados não estruturados).

Alguns aproveitam a análise integrada e a tecnologia de banco de dados in-memory (que mantém o conjunto de dados na memória do computador em vez de no armazenamento em disco) para fornecer acesso em tempo real a dados confiáveis e impulsionar a tomada de decisões.

Sem DW é muito difícil combinar dados de fontes heterogêneas, garantir que estejam no formato certo para análise e obter uma visão atual e de longo alcance dos dados ao longo do tempo.



Definindo Data Warehouses

Benefícios do DW:

- **Melhor Análise de Negócios:** com o DW, os tomadores de decisão têm acesso a dados de várias fontes e não precisam mais tomar decisões com base em informações incompletas.
- **Consultas Mais Rápidas:** os DWs são construídos especificamente para recuperação e análise rápida de dados. Com um DW, você pode consultar rapidamente grandes quantidades de dados consolidados com pouco ou nenhum suporte de TI.
- **Melhoria da Qualidade dos Dados:** antes de serem carregados no DW, os dados passam por um processo de limpeza garantindo que os dados sejam transformados em um formato consistente para apoiar análises - e decisões - com base em dados precisos e de alta qualidade.
- **Visão Histórica:** ao armazenar dados históricos ricos, um data warehouse permite que os tomadores de decisão aprendam com tendências e desafios passados, façam previsões e conduzam a melhoria contínua dos negócios.



Definindo Data Warehouses

Temos um curso inteiro aqui na DSA sobre como construir DWs locais e em nuvem.

É o primeiro curso da Formação Engenheiro de Dados, o curso de Design e Implementação de Data Warehouses.

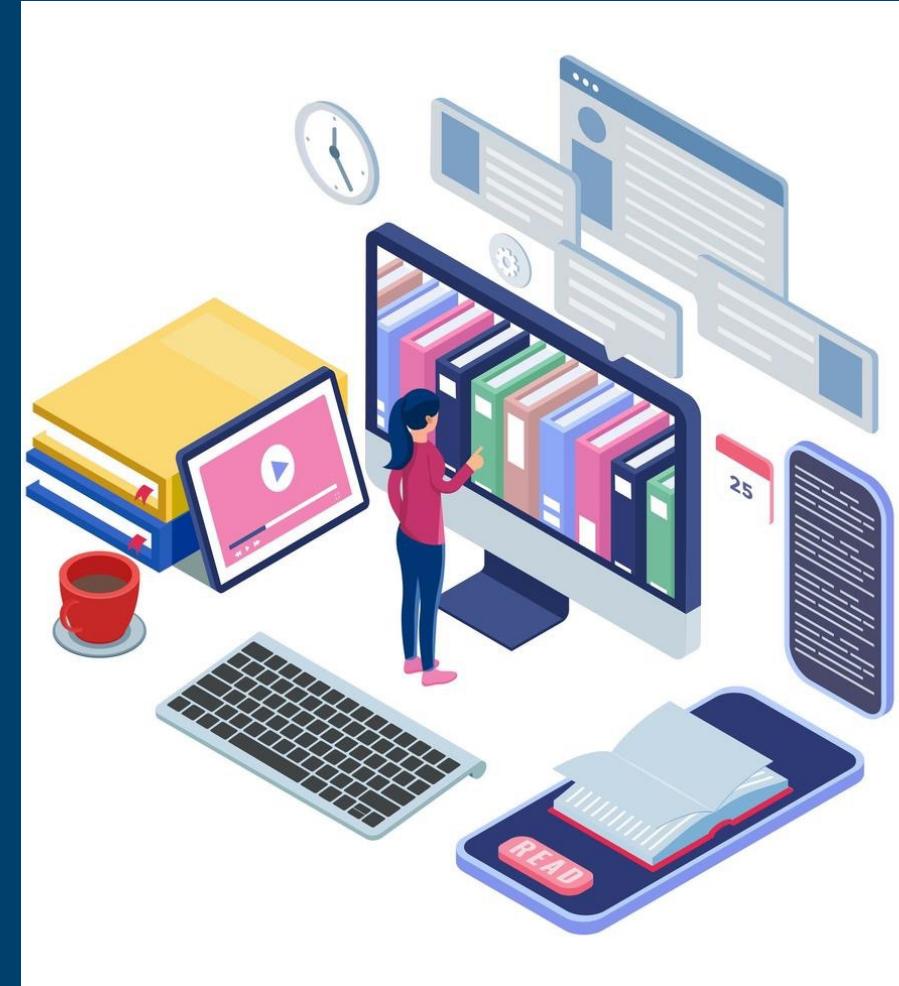


Big Data Fundamentos 3.0

Definindo Data Lakes

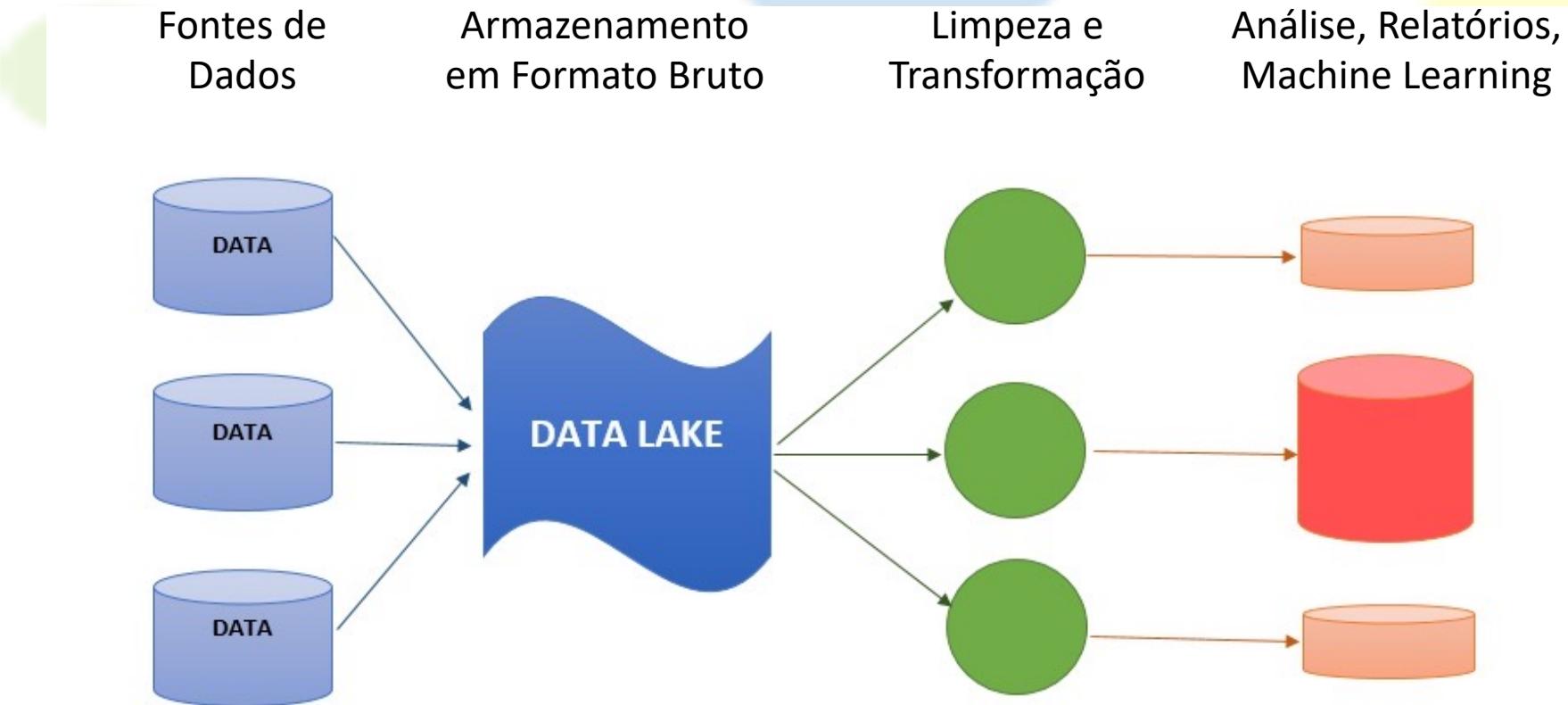
Data Science Academy

Data Science Academy





Definindo Data Lakes

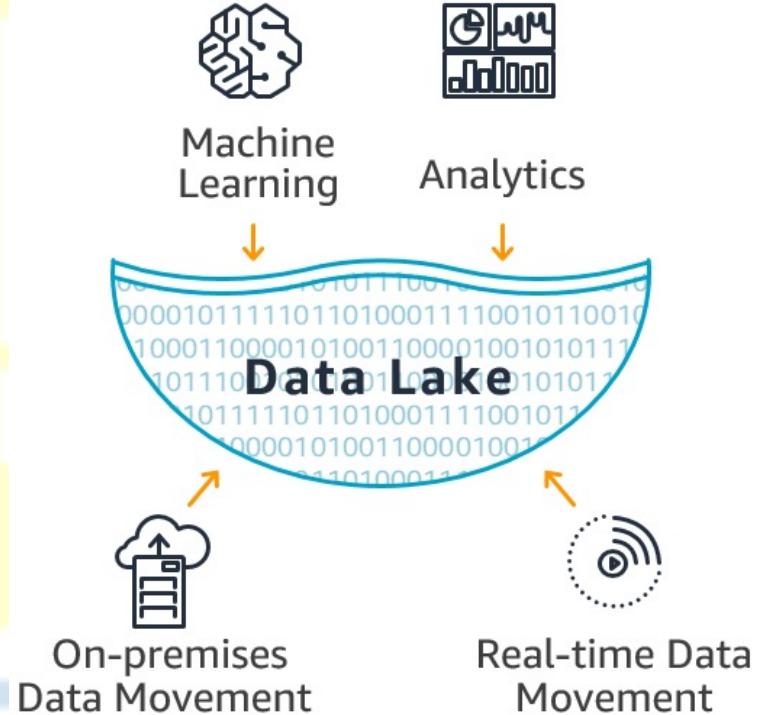


Definindo Data Lakes

Um Data Lake é um repositório centralizado que permite armazenar todos os dados estruturados e não estruturados em qualquer escala. Podemos armazenar os dados como estão na fonte, sem ter que primeiro estruturá-los e executar diferentes tipos de análises - de painéis e visualizações a processamento de Big Data, análises em tempo real e aprendizado de máquina para orientar melhores decisões.

Dependendo dos requisitos, uma empresa típica exigirá um Data Warehouse e um Data Lake, pois eles atendem a diferentes necessidades e casos de uso.

A estrutura dos dados ou schema (esquema) não é definida quando os dados são capturados. Isso significa que você pode armazenar todos os dados em formato bruto sem a necessidade de saber quais perguntas de negócio deverão ser respondidas no futuro.

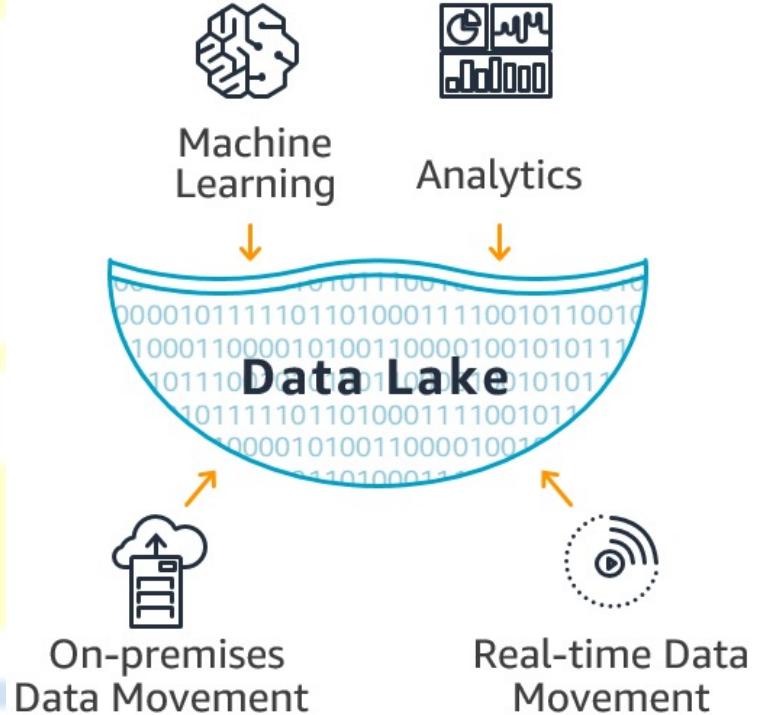


Definindo Data Lakes

Diferentes tipos de análises, como consultas SQL, análises de Big Data, pesquisa de texto, análises em tempo real e aprendizado de máquina, podem ser usados para descobrir insights.

Os Data Lakes permitem que as empresas gerem diferentes tipos de percepções sobre os dados, desde relatórios sobre dados históricos até modelos preditivos criados com Machine Learning.

O principal desafio de uma arquitetura de Data Lake é que os dados brutos são armazenados sem supervisão do conteúdo. Para que um Data Lake torne os dados utilizáveis, ele precisa ter mecanismos definidos para catalogar e proteger os dados. Sem esses elementos, os dados não podem ser encontrados ou confiáveis, resultando em um “Pântano de Dados” (Data Swamp). Atender às necessidades de públicos mais amplos exige que os Data Lakes tenham governança, gestão de metadados, consistência semântica e controles de acesso.



Definindo Data Lakes

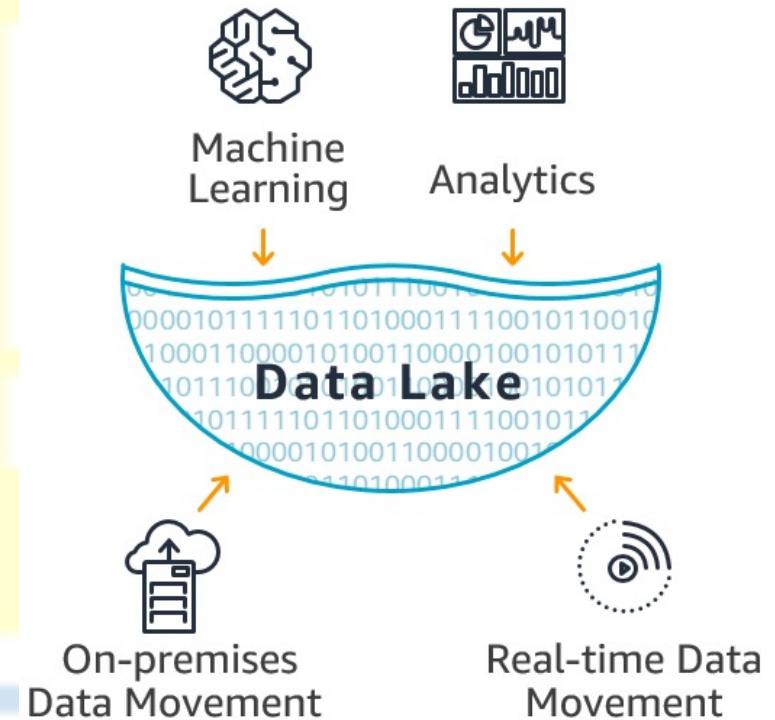
Data Lake é um conceito e pode ser construído com diferentes tecnologias como Apache Hadoop ou Bancos de Dados NoSQL.

Podemos importar dados do DW para o Data Lake e vice-versa dependendo das necessidades de negócio da empresa.

Para o DW normalmente usamos **ETL** (Extração, Transformação e Carga).

Para o Data Lake normalmente usamos **ELT** (Extração, Carga e Transformação).

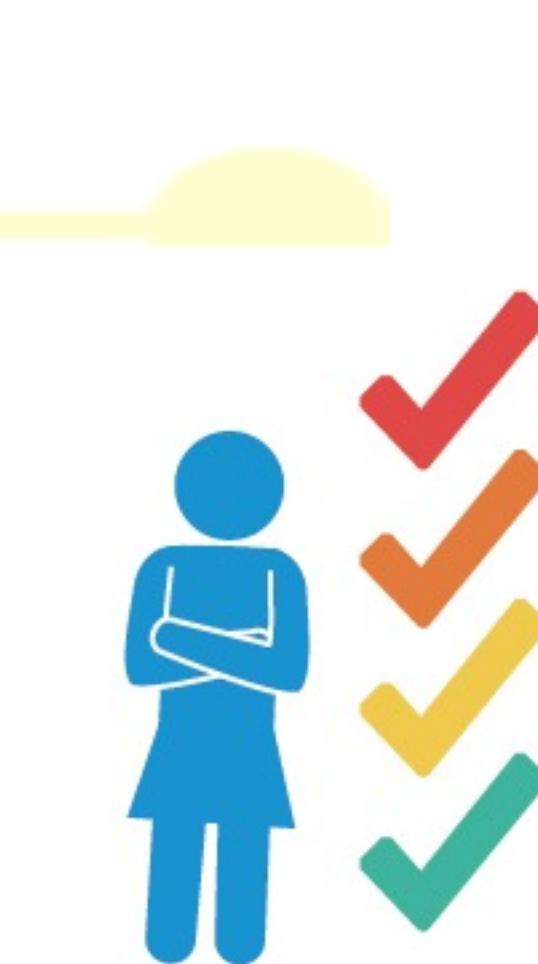
Data Lakes e DWs podem fazer parte de uma grande estrutura central de armazenamento chamada Data Hub.



Definindo Data Lakes

Benefícios do Data Lake:

- **Armazenamento em Formato Bruto:** não precisamos limpar e transformar os dados antes do armazenamento.
- **Importação de Qualquer Quantidade de Dados em Tempo Real:** os dados são coletados de várias fontes e movidos para o Data Lake em seu formato original. Este processo permite dimensionar dados de qualquer tamanho, enquanto economiza tempo de definição de estruturas de dados, esquema e transformações.
- **Repositório Central Para Todos os Dados da Empresa:** os Data Lakes permitem que várias funções como Cientistas de Dados, Engenheiros de Machine Learning, Analistas de Dados e Analistas de Negócios, acessem dados com sua ferramenta analítica específica.
- **Sem Necessidade de Movimentação dos Dados:** análises podem ser executadas sem necessidade de mover os dados para um sistema de análise separado.



Definindo Data Lakes

Temos um curso inteiro aqui na DSA sobre como construir Data Lakes locais e em nuvem.

É o segundo curso da Formação Engenheiro de Dados, o curso Data Lake – Design, Projeto e Integração.

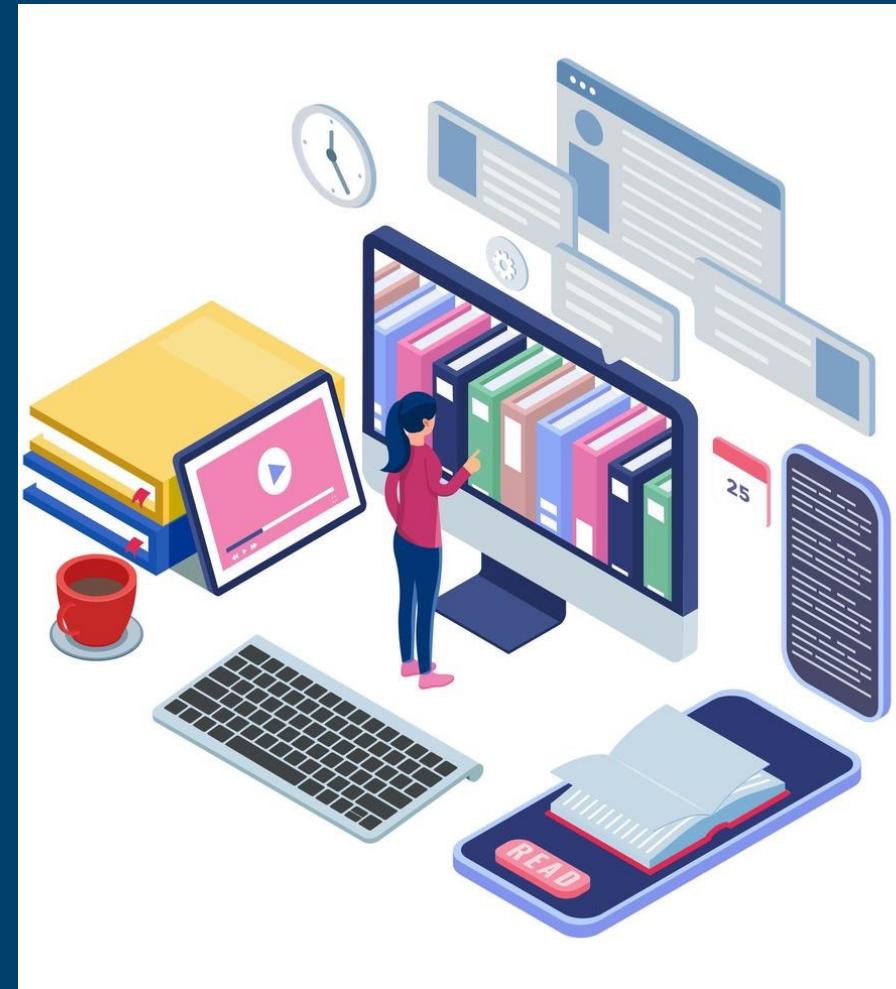


Big Data Fundamentos 3.0

Definindo Data Stores

Data Science Academy

Data Science Academy





Definindo Data Stores

Um Data Store é um repositório para armazenar e gerenciar de forma persistente coleções de dados que incluem não apenas dados estruturados, mas também tipos de armazenamento variado, como documentos, dados no formato de chave-valor, filas de mensagens e outros formatos de arquivo.

Os tipos mais comuns de Data Stores:

- Armazenamento de chave-valor (Redis, Memcached)
- Motor de pesquisa de texto completo (Elastic Search)
- Fila de mensagens (Apache Kafka)
- Sistema de arquivos distribuídos (Hadoop HDFS, AWS S3)



Definindo Data Stores

Benefícios do Data Store:

- **Armazenamento de Variados Tipos de Dados:** dados que não se encaixam em outros repositórios de armazenamento.
- **Flexibilidade:** armazenamento de dados aderente às necessidades da aplicação final.
- **Suporte a Dados Semi-Estruturados:** dados que possuem alguma organização prévia, mas que devem ser usados em seu formato original.
- **Custo Total Menor:** por se tratar de um tipo simplificado de armazenamento o custo total tende a ser menor que outra solução de armazenamento.



Definindo Data Stores

Temos um curso inteiro aqui na DSA voltado para Modelagem de diferentes sistemas de armazenamento.

É o segundo curso da Formação Arquiteto de Dados, o curso Modelagem de Bancos de Dados Relacionais, Não Relacionais e Data Stores.



Big Data

Fundamentos 3.0

Sistemas Híbridos de Armazenamento

Data Science Academy

Data Science Academy





Sistemas Híbridos de Armazenamento

Com o avanço do Big Data veremos cada vez mais sistemas híbridos de armazenamento, com dados armazenados em diferentes tipos de repositórios, local ou na nuvem.





Sistemas Híbridos de Armazenamento

DWs, Data Lakes e Data Stores serão usados em conjunto criando assim uma grande estrutura de armazenamento de dados, um Data Hub.



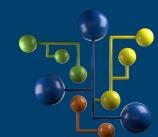
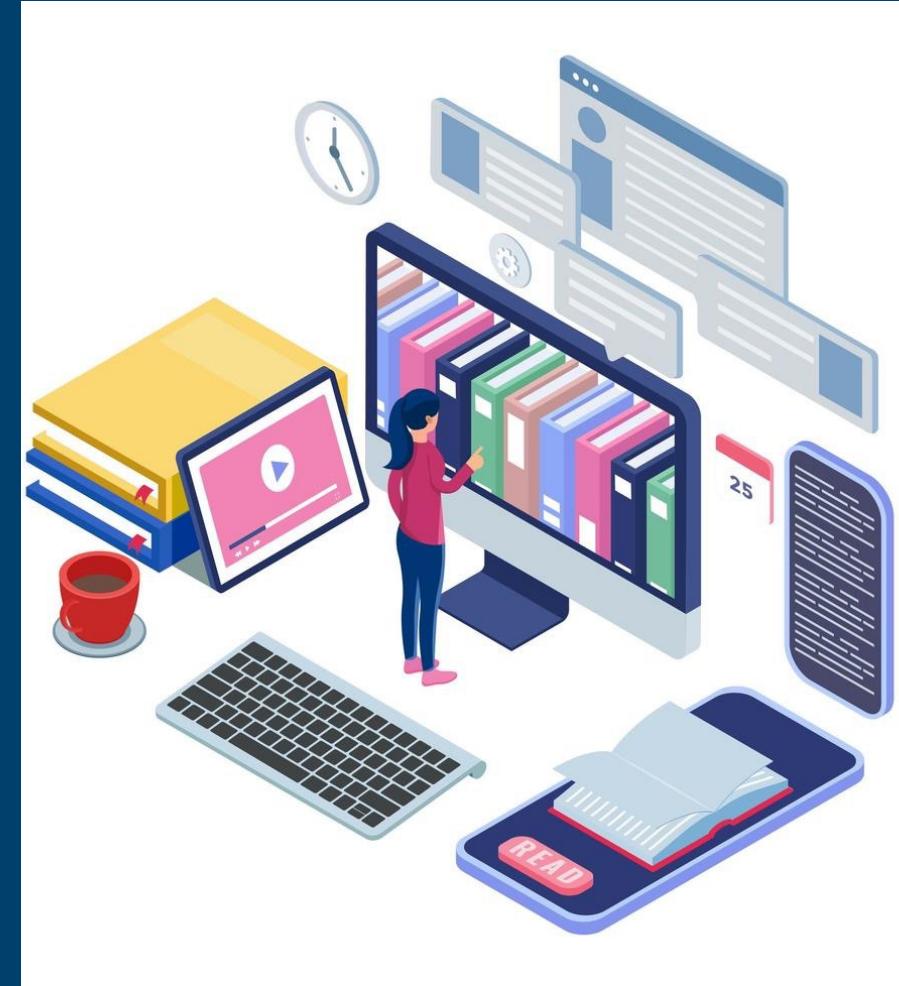
Big Data

Fundamentos 3.0

Armazenamento e Processamento Paralelo

Data Science Academy

Data Science Academy



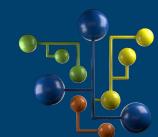
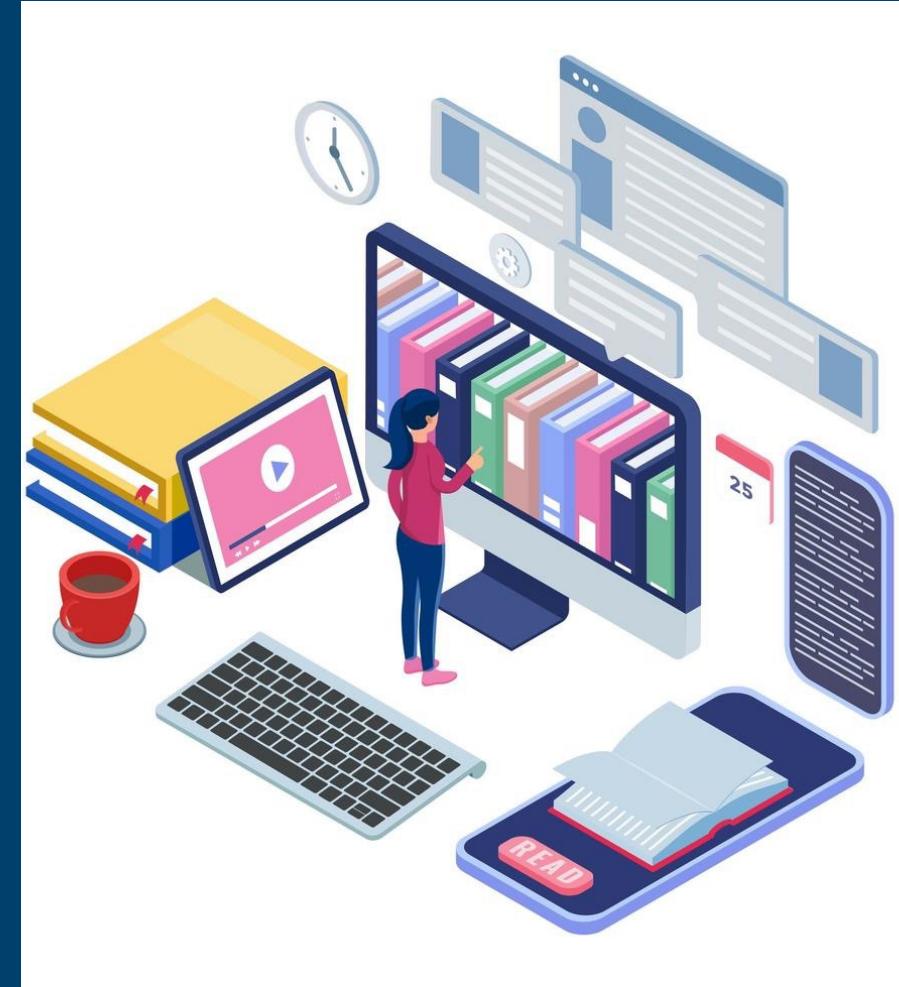
Big Data

Fundamentos 3.0

O Que é um Cluster de Computadores?

Data Science Academy

Data Science Academy





O Que é um Cluster de Computadores?



Um servidor é um computador, geralmente com alta capacidade computacional, que “serve” (fornecer) serviços de armazenamento, aplicações ou bancos de dados.





O Que é um Cluster de Computadores?



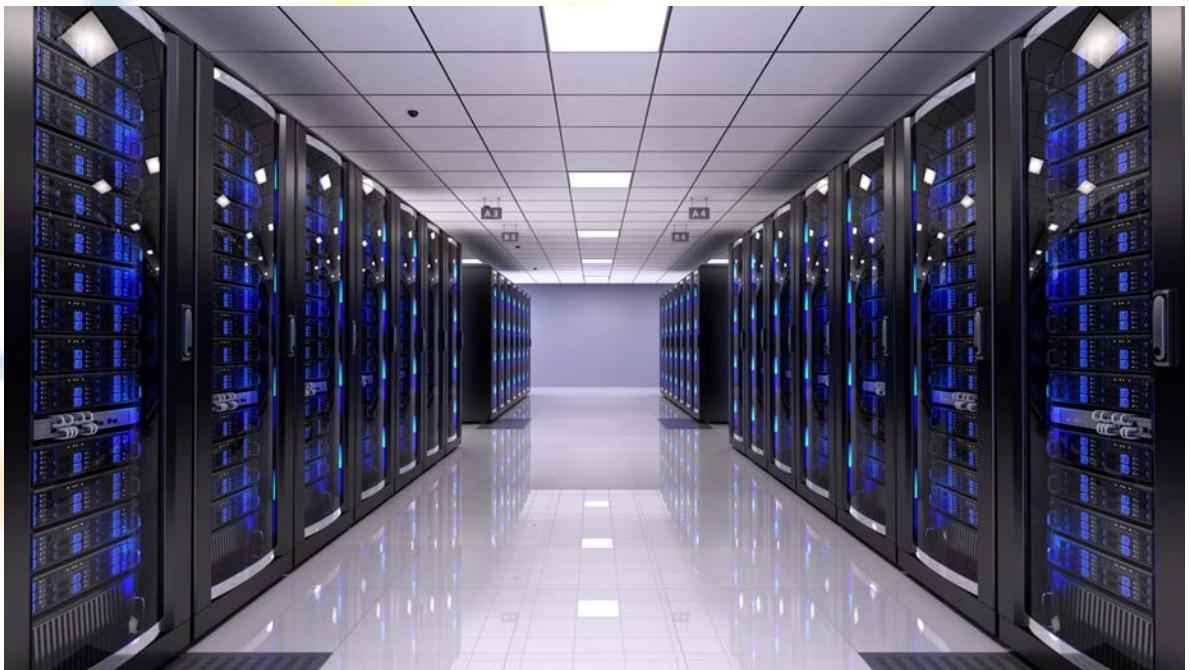
Um servidor possui escalabilidade vertical, ou seja, há um limite até onde conseguimos incluir mais espaço em disco, mais processadores e mais memória RAM.





O Que é um Cluster de Computadores?

Um cluster de computadores é um conjunto de servidores com um mesmo propósito visando fornecer um tipo de serviço, como armazenamento ou processamento de dados.





O Que é um Cluster de Computadores?

Um cluster possui escalabilidade horizontal, ou seja, se quisermos aumentar a capacidade computacional incluímos mais máquinas no cluster (além da escalabilidade vertical de cada máquina individual no cluster).





O Que é um Cluster de Computadores?

Clusters de computadores são cada vez mais usados em Big Data, o que nos permite realizar armazenamento e processamento paralelo através de diversas máquinas (diversos servidores).

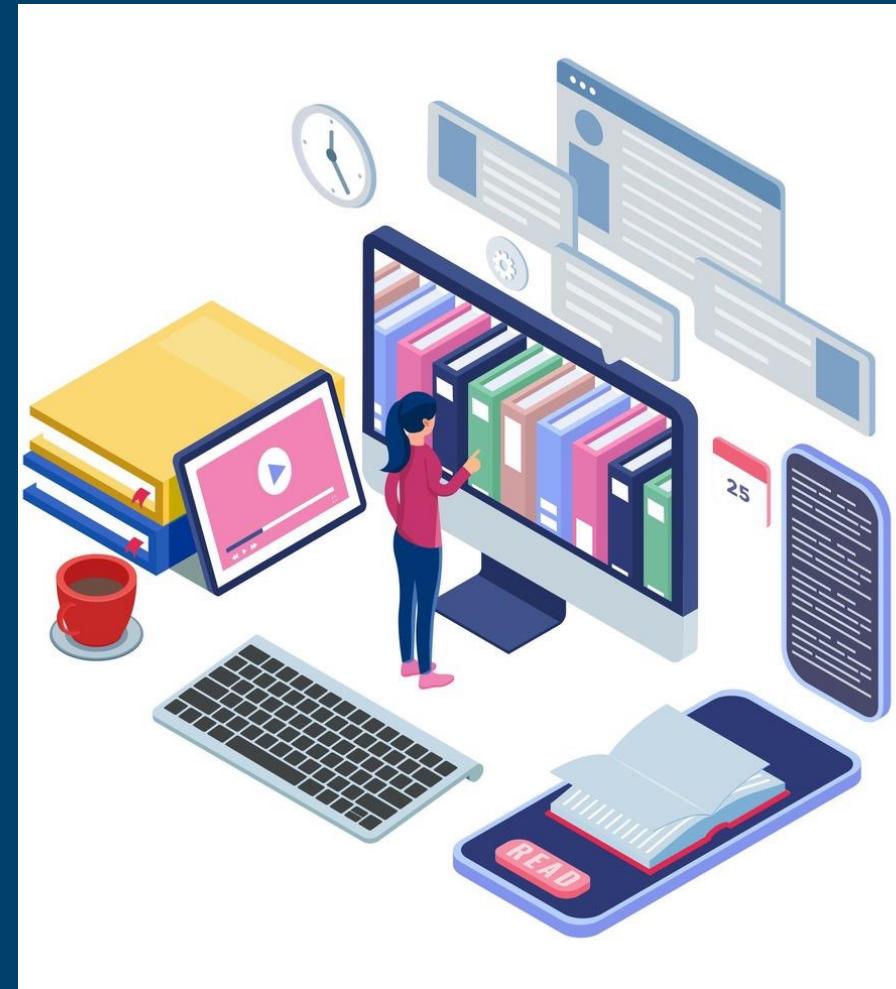


Big Data Fundamentos 3.0

O Que é Armazenamento Paralelo?

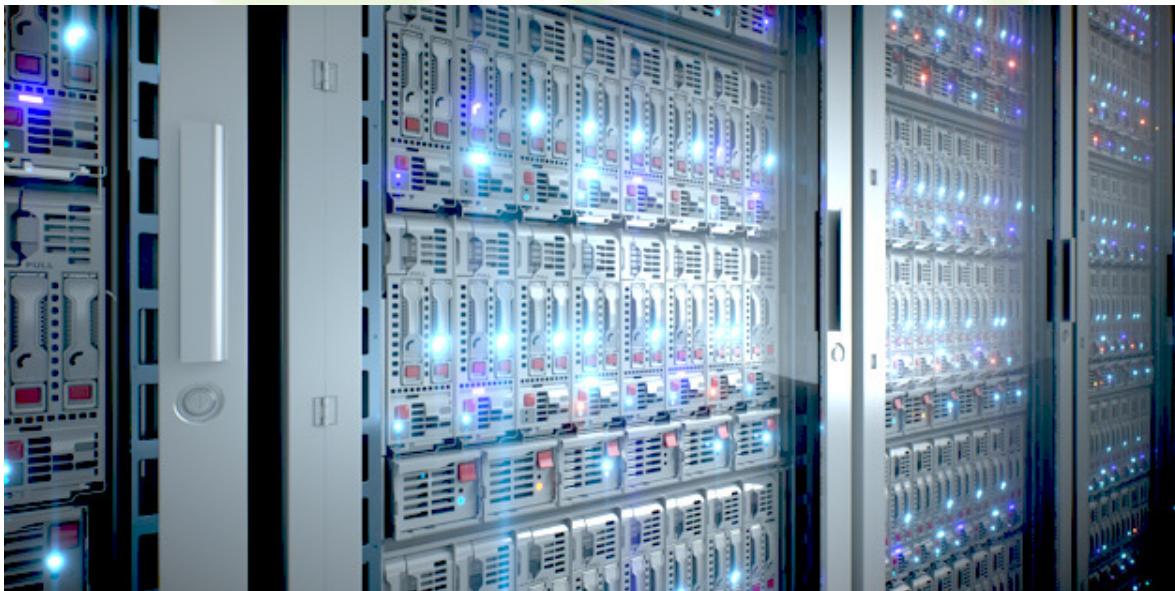
Data Science Academy

Data Science Academy





O Que é Armazenamento Paralelo?



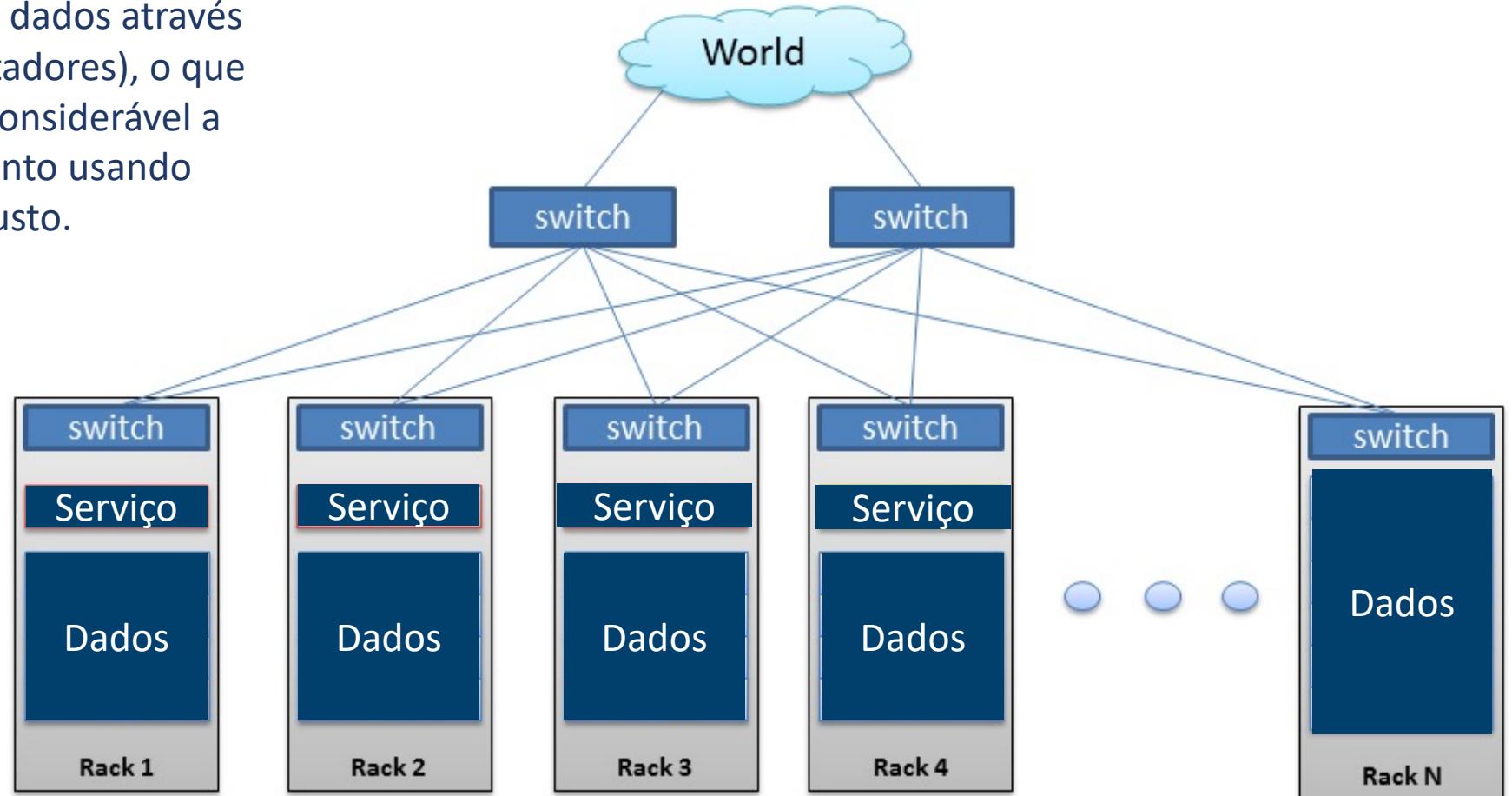
Com clusters de computadores aumentamos de forma considerável a capacidade computacional.





O Que é Armazenamento Paralelo?

O armazenamento paralelo consiste em distribuir o armazenamento de dados através de diversos servidores (computadores), o que permite aumentar de forma considerável a capacidade de armazenamento usando hardware de baixo custo.

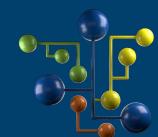


Big Data Fundamentos 3.0

Software para Armazenamento Paralelo - Apache Hadoop

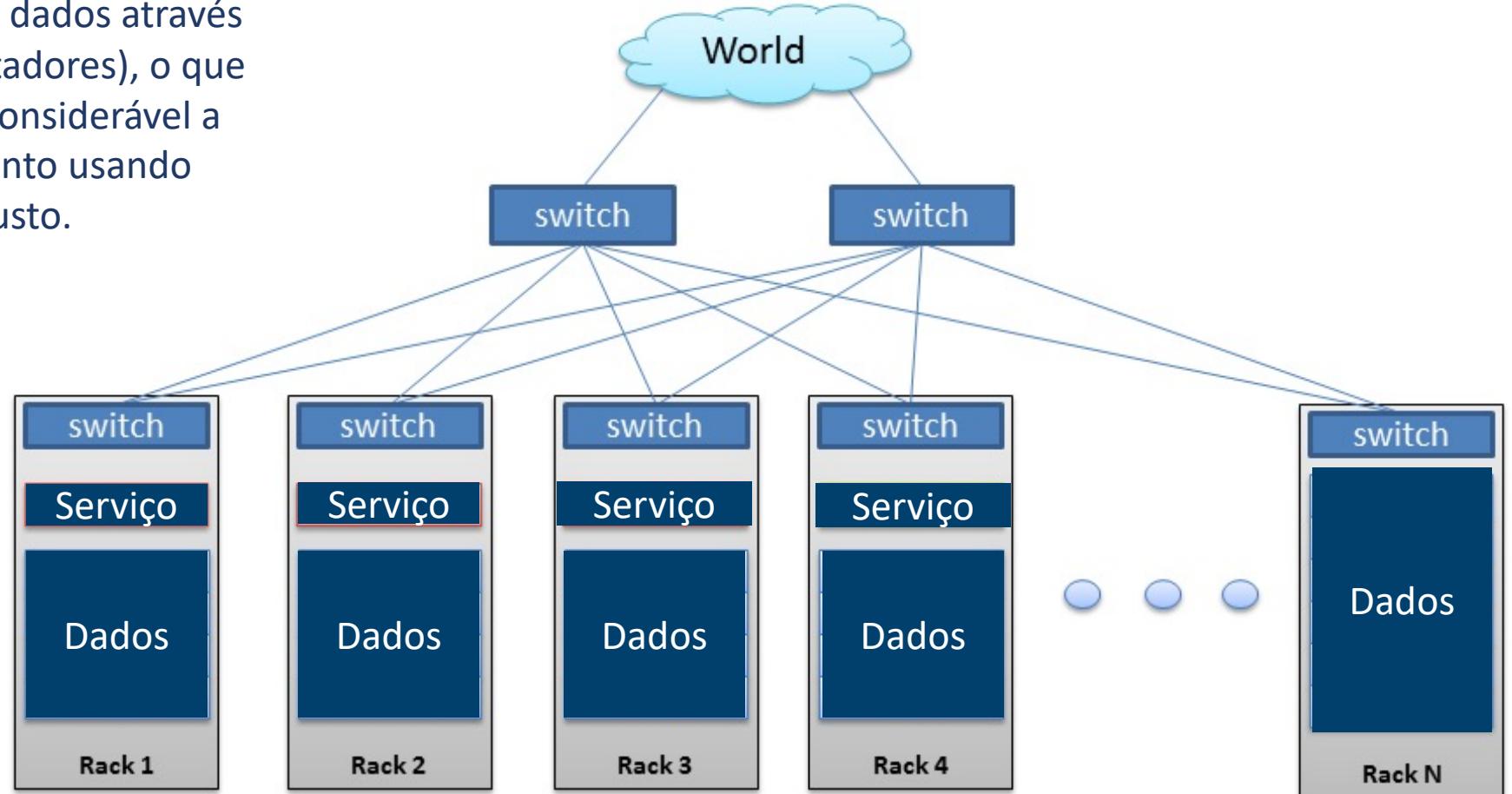
Data Science Academy

Data Science Academy



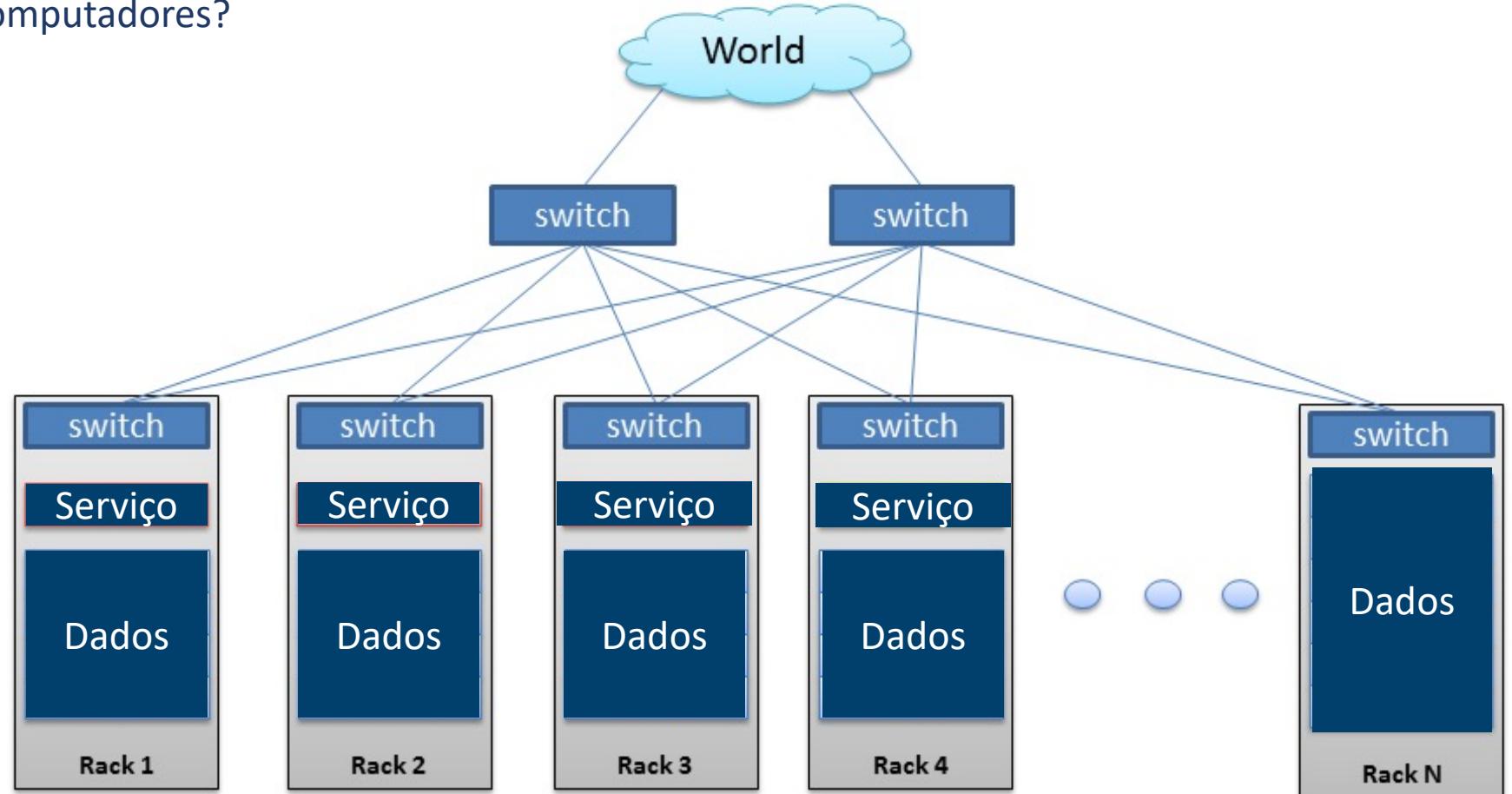
Software para Armazenamento Paralelo - Apache Hadoop

O armazenamento paralelo consiste em distribuir o armazenamento de dados através de diversos servidores (computadores), o que permite aumentar de forma considerável a capacidade de armazenamento usando hardware de baixo custo.



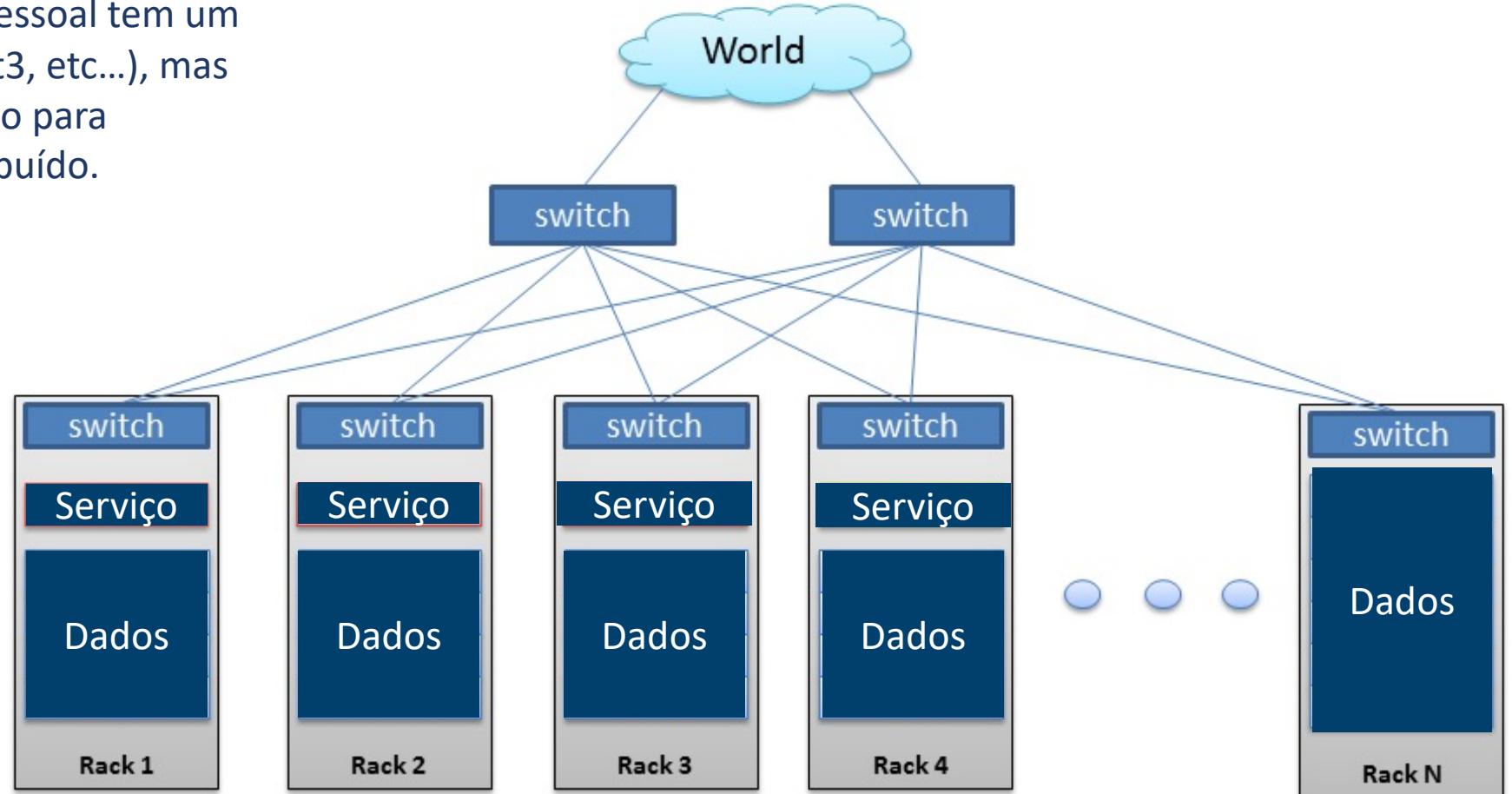
Software para Armazenamento Paralelo - Apache Hadoop

E como gerenciamos o armazenamento paralelo através de diversos computadores?



Software para Armazenamento Paralelo - Apache Hadoop

Precisamos de um sistema de arquivos distribuído. Seu computador pessoal tem um sistema de arquivos (NTFS, ext3, etc...), mas ele não foi desenvolvido para armazenamento distribuído.



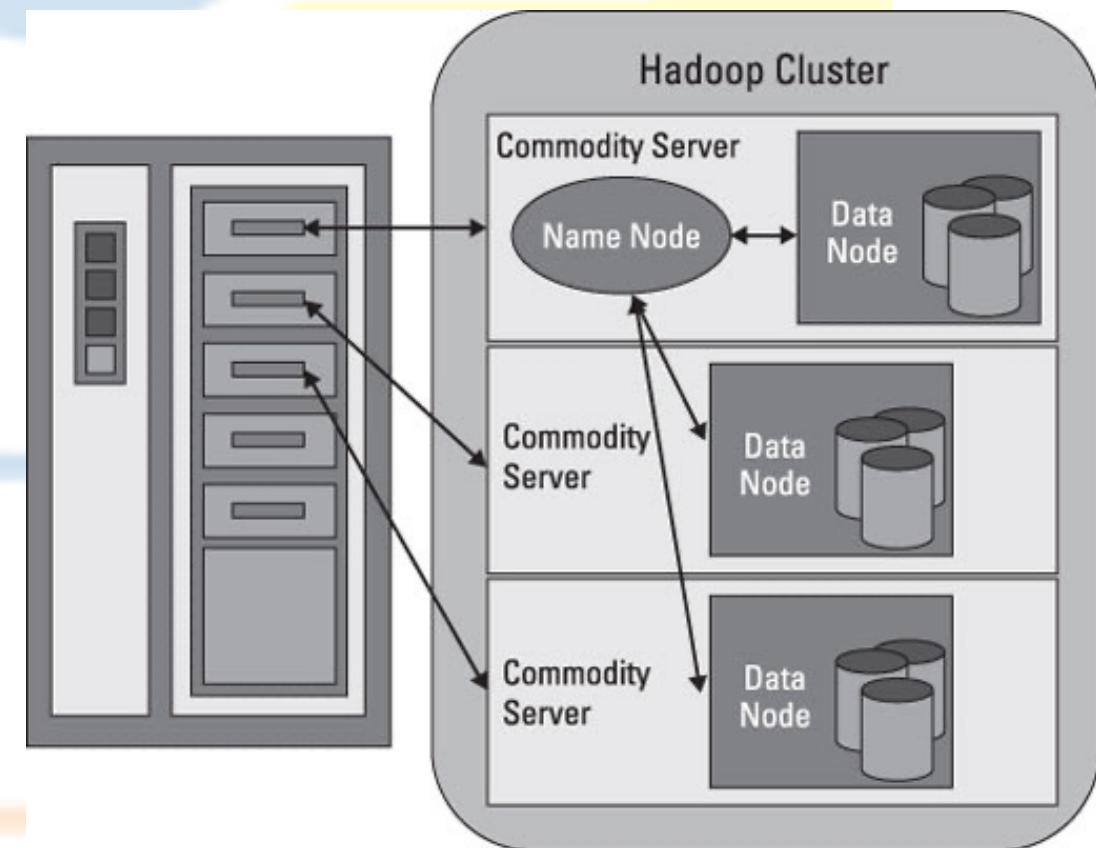
Software para Armazenamento Paralelo - Apache Hadoop

Entre algumas opções, o Apache Hadoop HDFS (Hadoop Distributed File System) tem se mostrado a solução ideal para gerenciar o armazenamento distribuído em um cluster de computadores.

O HDFS é o software responsável pela gestão do cluster de computadores definindo como os arquivos serão distribuídos através do cluster.

Com o HDFS podemos construir um Data Lake que roda sobre um cluster de computadores e permite o armazenamento de grandes volumes de dados com hardware commodity (de baixo custo).

Isso permitiu que o Big Data pudesse ser usado em larga escala!



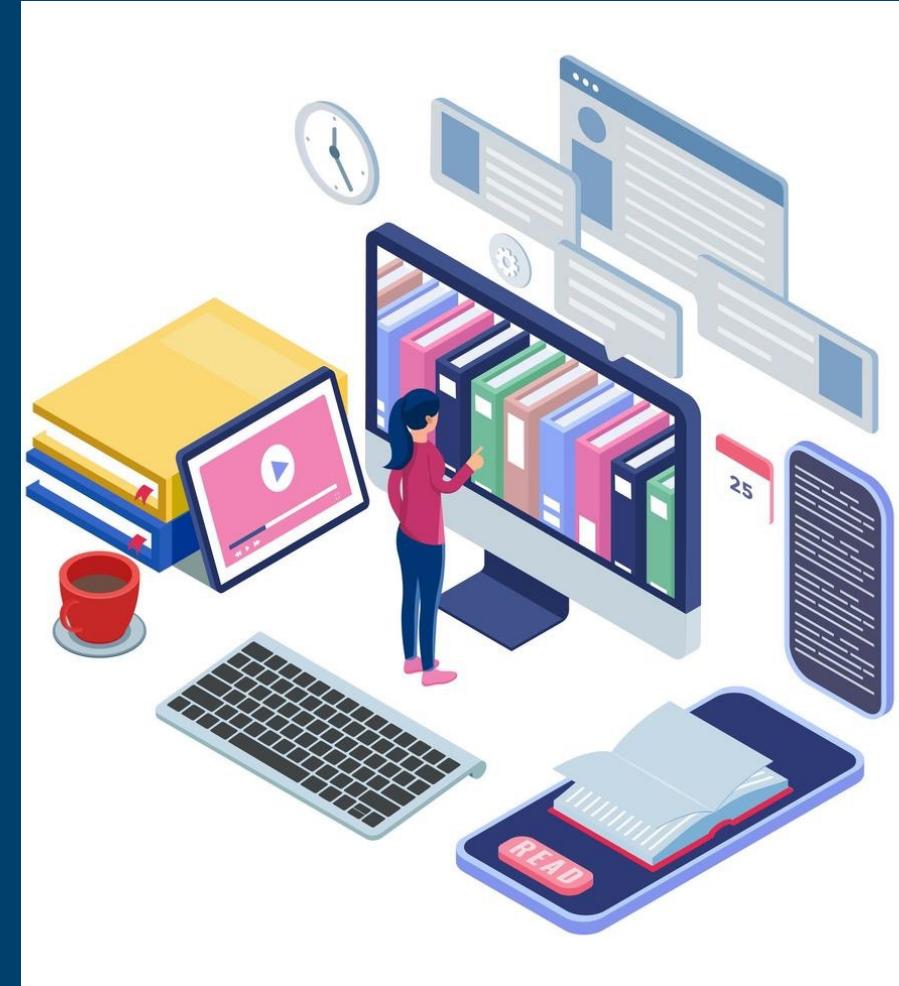
Big Data

Fundamentos 3.0

Processamento Paralelo de Big Data

Data Science Academy

Data Science Academy





Processamento Paralelo de Big Data

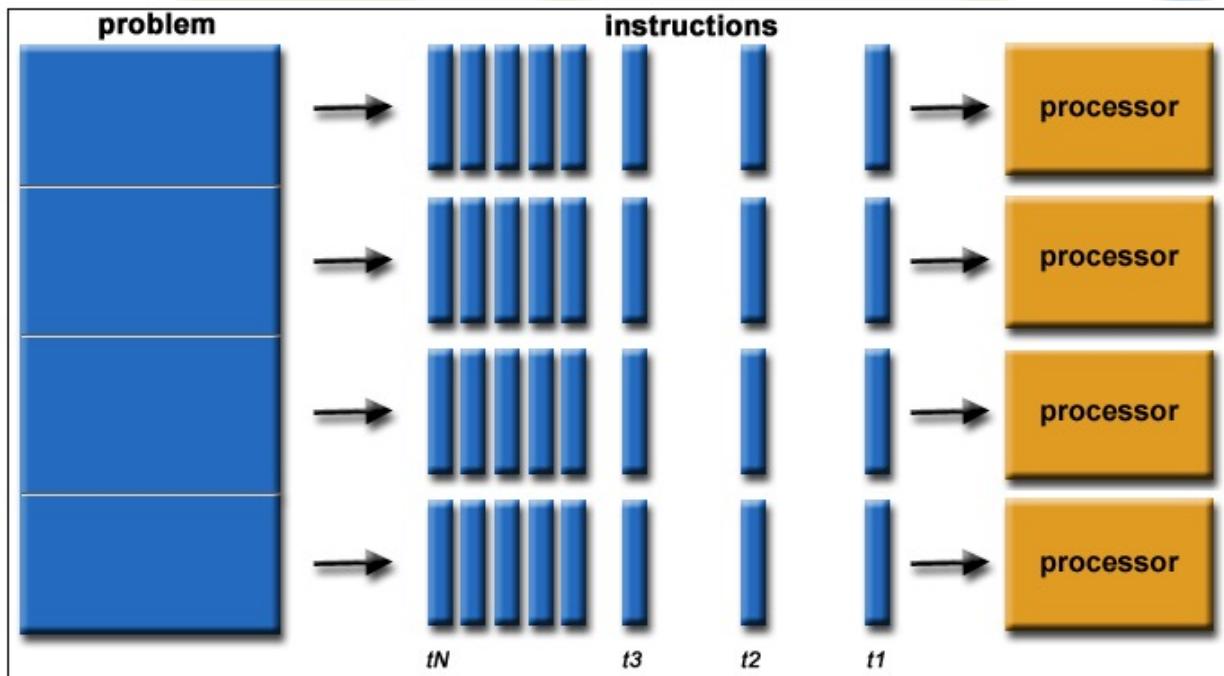
Resolvemos um problema! Podemos agora armazenar grandes quantidades de dados em um cluster de computadores através de armazenamento paralelo de dados.

Mas como vamos processar os dados se eles estão agora distribuídos em diversos computadores?





Processamento Paralelo de Big Data



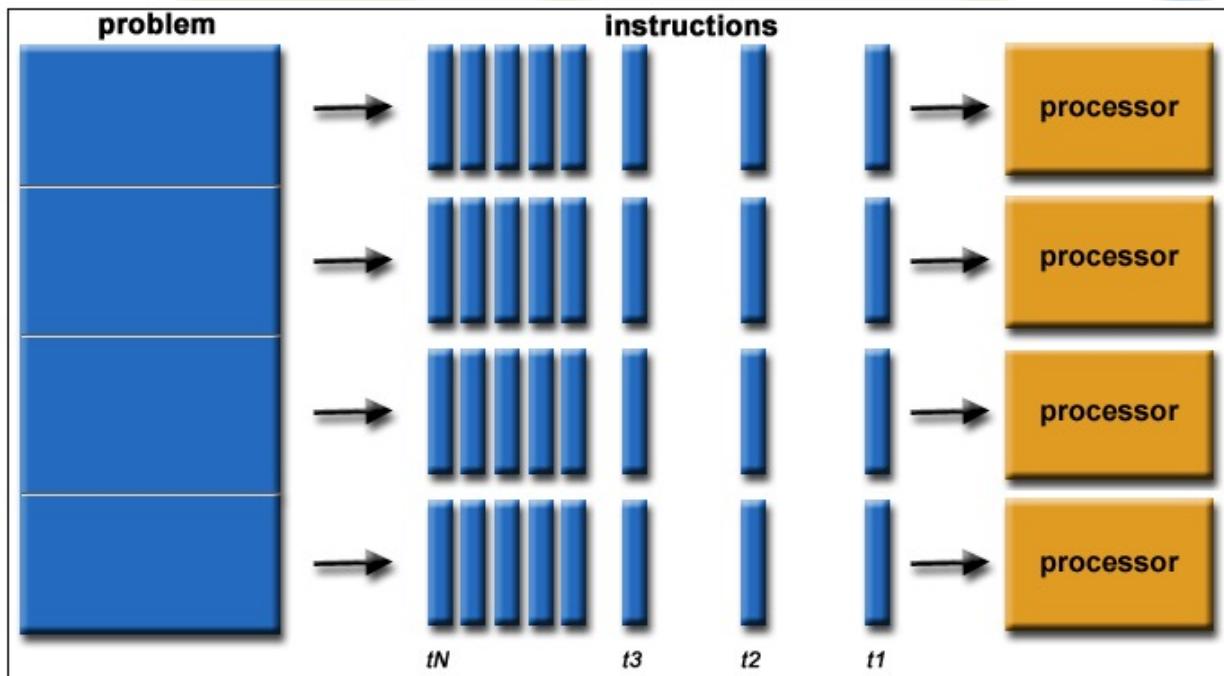
No processamento paralelo o objetivo é dividir uma tarefa em várias sub-tarefas e executá-las em paralelo.

O Apache Hadoop MapReduce e o Apache Spark são dois frameworks para esse propósito.





Processamento Paralelo de Big Data



Ao usar um framework de processamento paralelo, as sub-tarefas são levadas para o processador da máquina do cluster onde os dados estão armazenados, aumentando assim a velocidade de processamento de grandes volumes de dados.



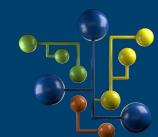
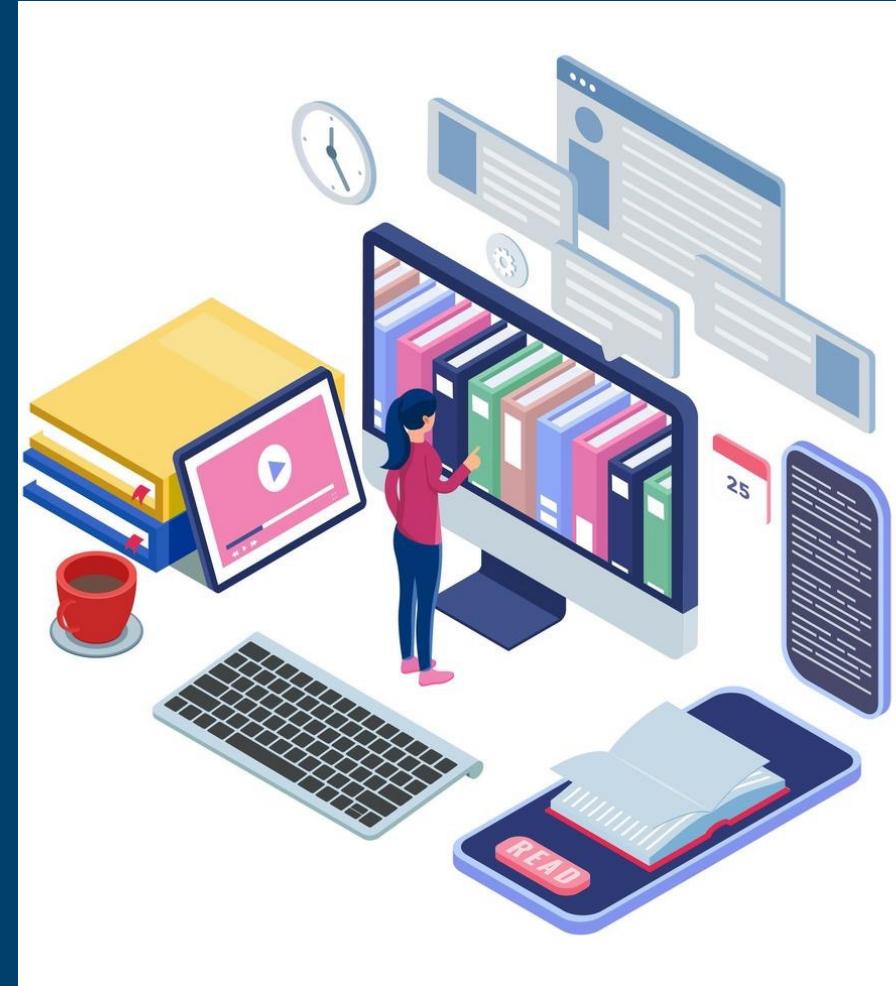
Big Data

Fundamentos 3.0

Arquitetura de Armazenamento e Processamento Paralelo

Data Science Academy

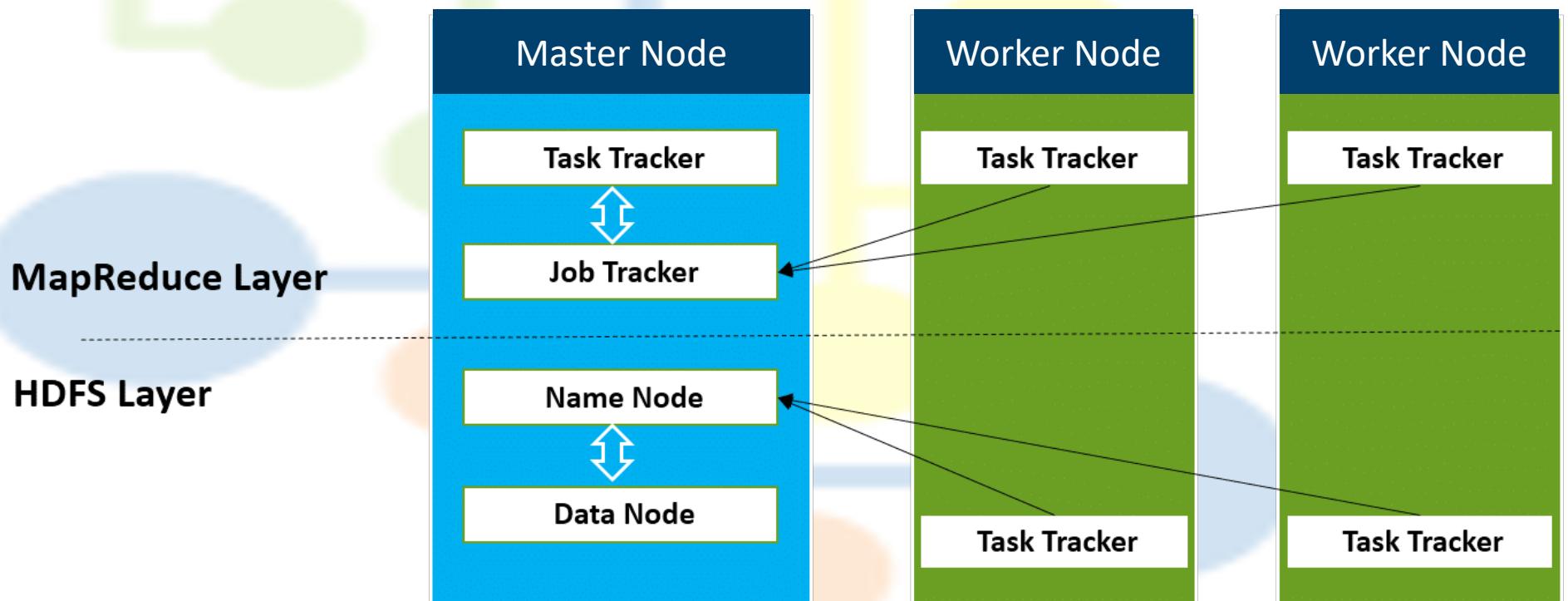
Data Science Academy





Arquitetura de Armazenamento e Processamento Paralelo

Considerando o Apache Hadoop, teríamos o seguinte esquema:



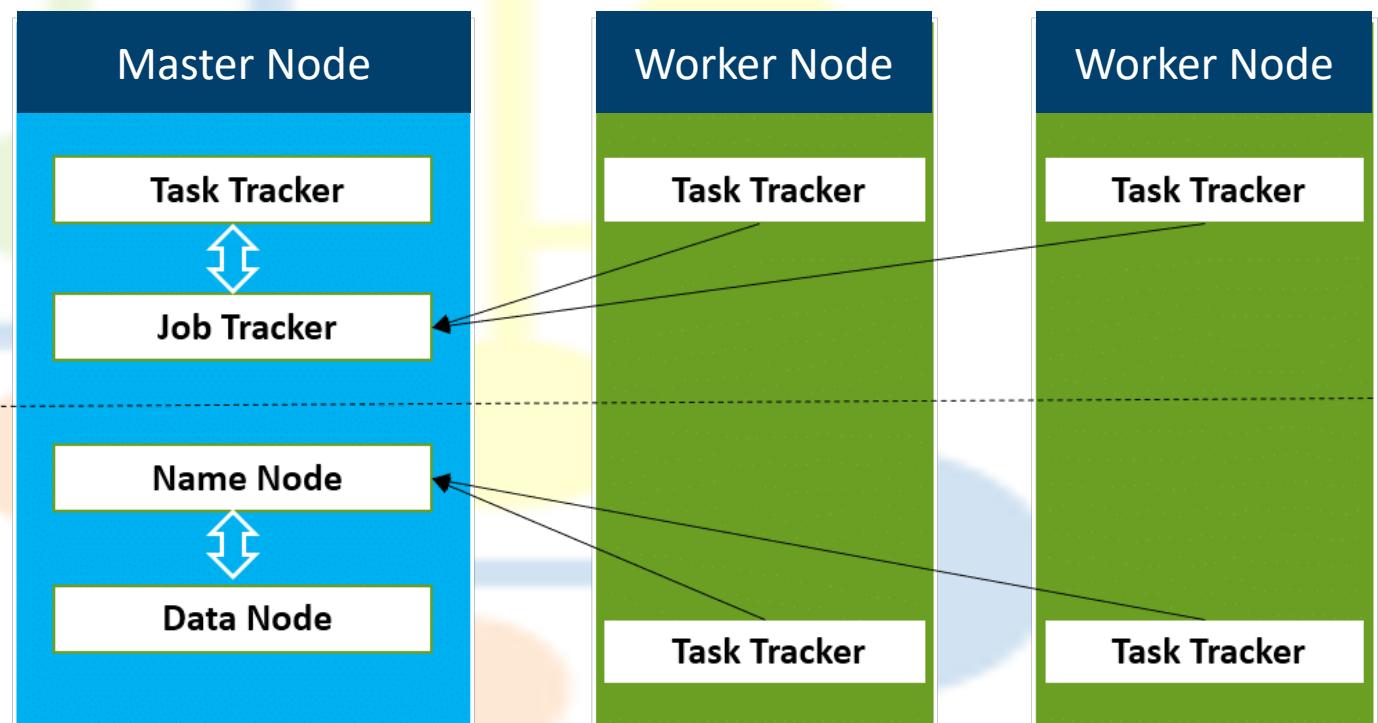


Arquitetura de Armazenamento e Processamento Paralelo

O HDFS é um serviço rodando em todas as máquinas do cluster, sendo um NameNode para gerenciar o cluster e os DataNodes que fazem o trabalho de armazenamento propriamente dito.

MapReduce Layer

HDFS Layer



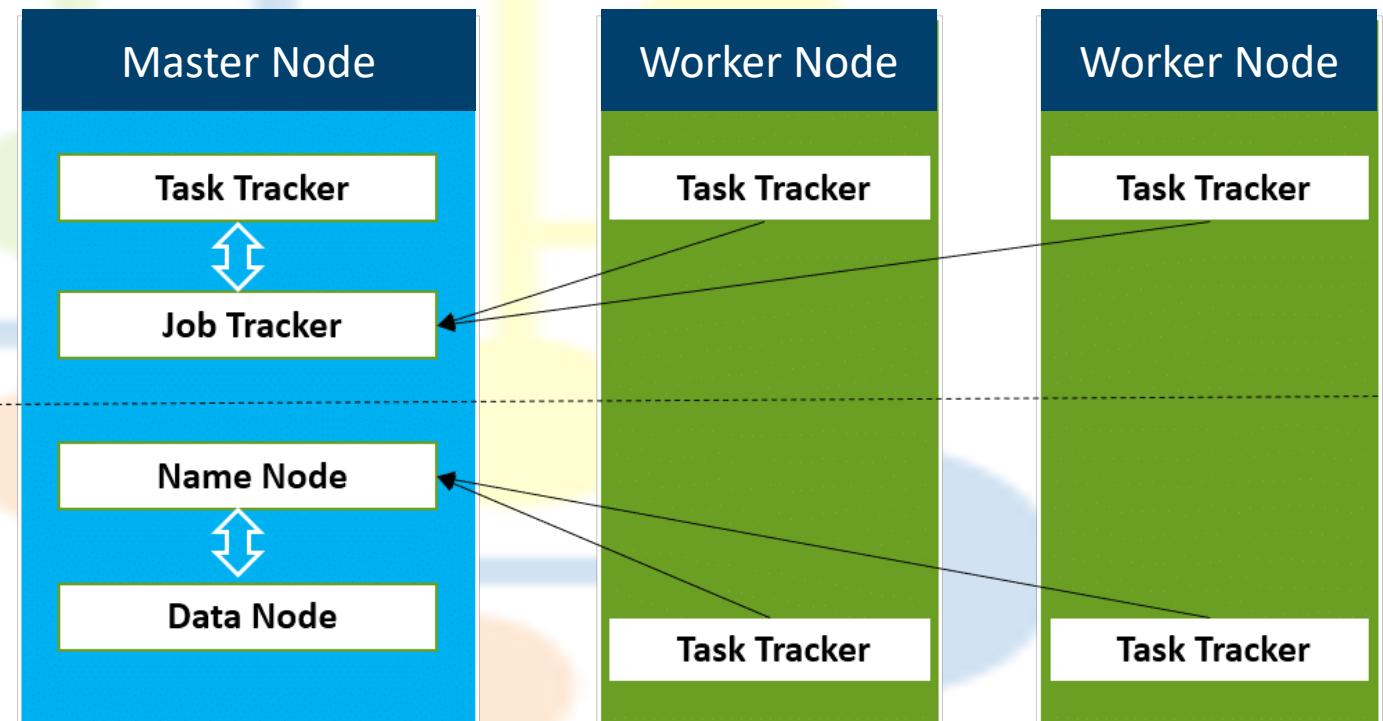


Arquitetura de Armazenamento e Processamento Paralelo

O MapReduce também é um serviço rodando em todas as máquinas do cluster, sendo um Job Tracker para gerenciar o processamento e os Task Trackers que fazem o trabalho de processamento.

MapReduce Layer

HDFS Layer



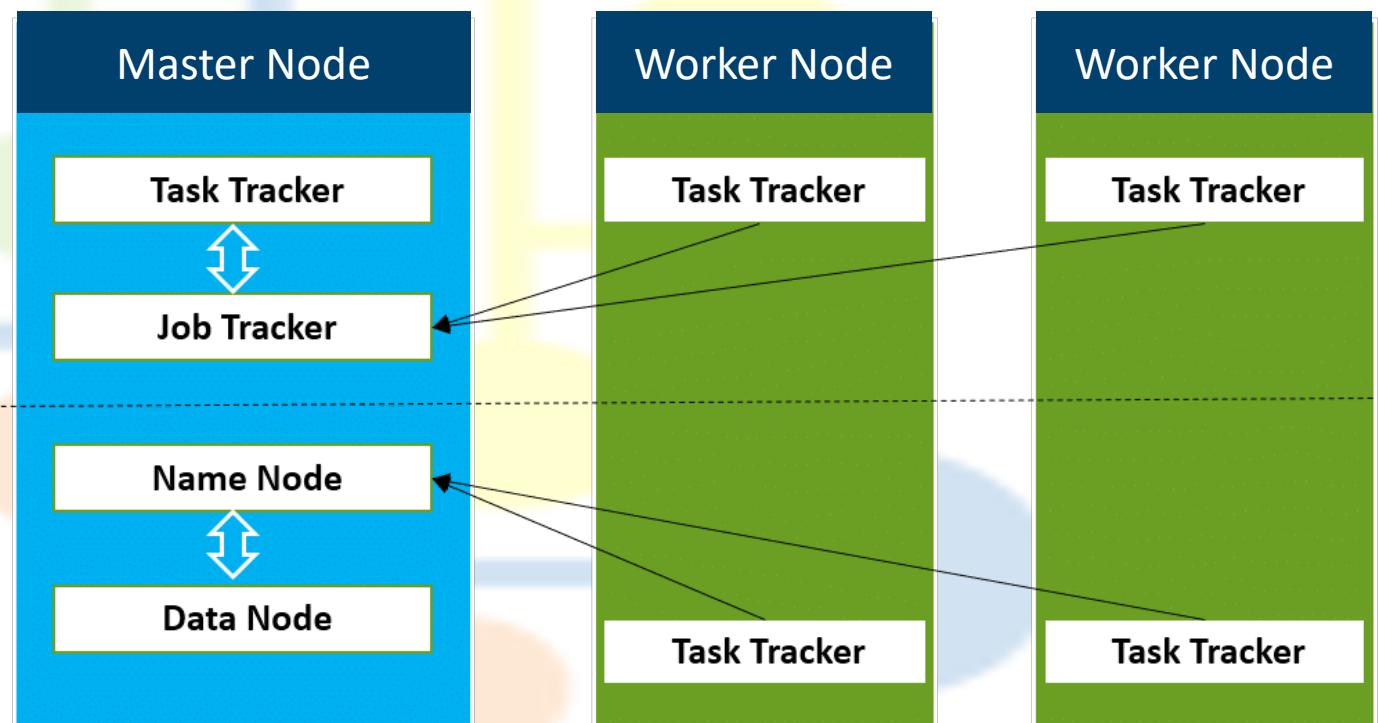


Arquitetura de Armazenamento e Processamento Paralelo

O Job Tracker consulta o NameNode a fim de saber a localização dos blocos de dados nas máquinas do cluster.

MapReduce Layer

HDFS Layer



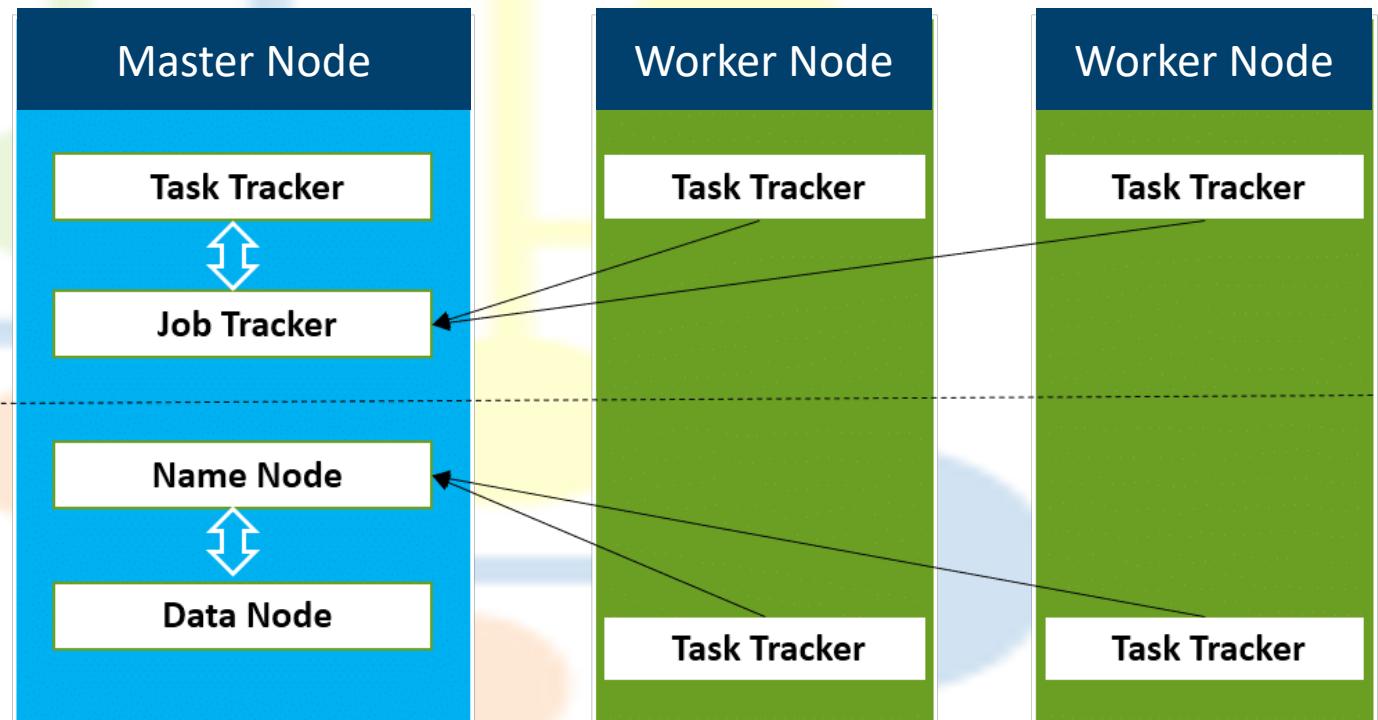


Arquitetura de Armazenamento e Processamento Paralelo

Os Task Trackers se comunicam com os DataNodes para obter os dados do disco, executar o processamento e então retornar o resultado ao Job Tracker.

MapReduce Layer

HDFS Layer



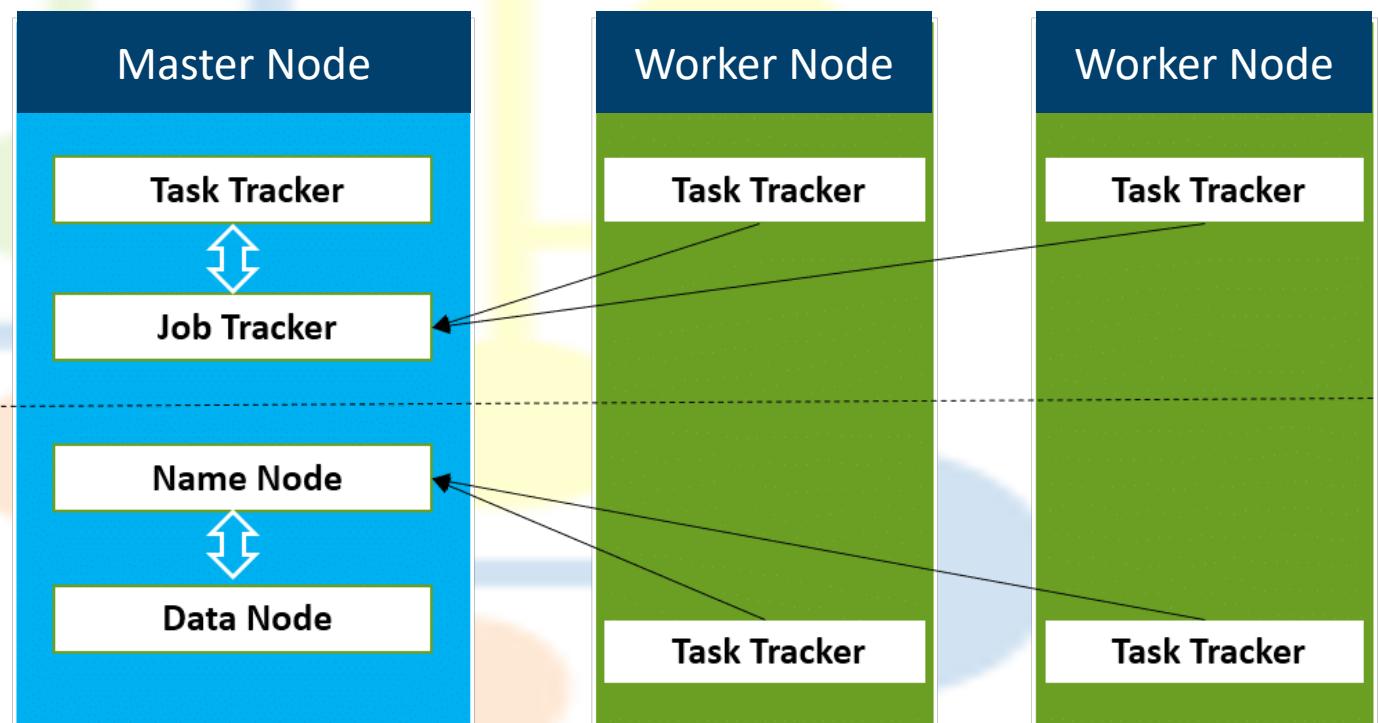


Arquitetura de Armazenamento e Processamento Paralelo

Essa arquitetura permite armazenar e processar grandes quantidades de dados e assim extrair valor do Big Data através da análise de dados.

MapReduce Layer

HDFS Layer



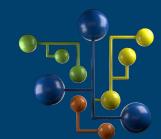
Big Data

Fundamentos 3.0

Cloud Computing

Data Science Academy

Data Science Academy

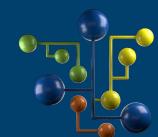
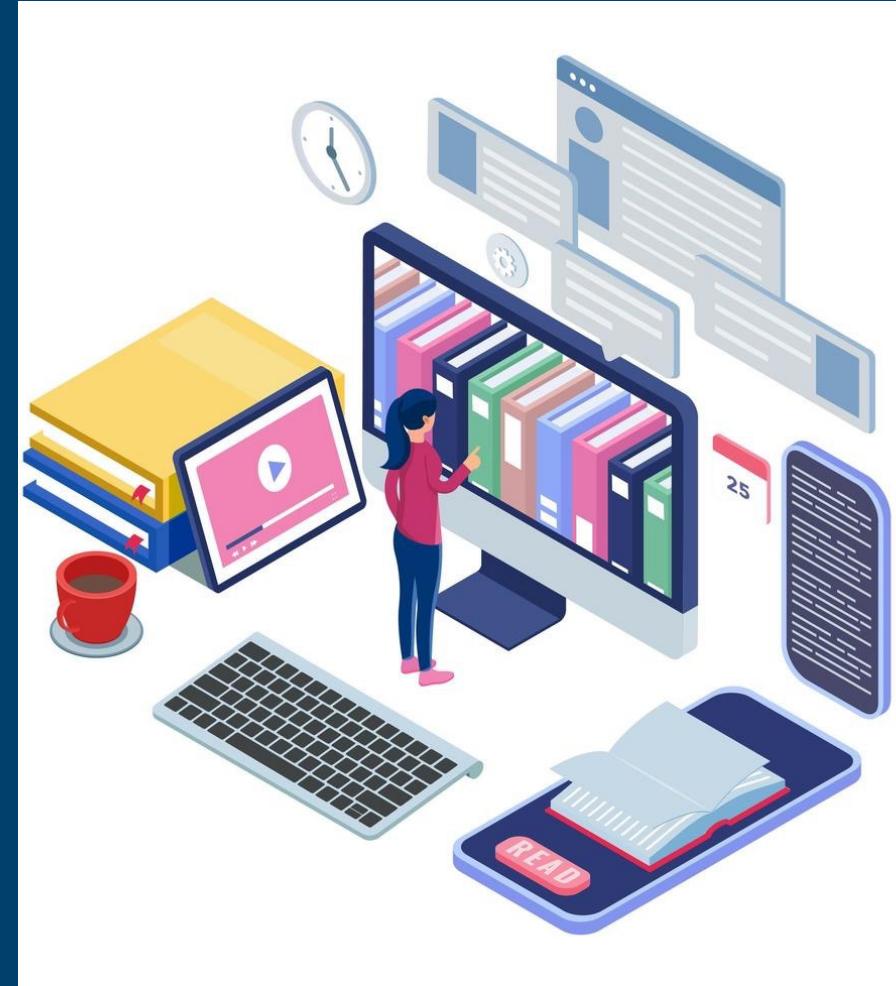


Big Data Fundamentos 3.0

O Que é Cloud Computing?

Data Science Academy

Data Science Academy





O Que é Cloud Computing?





O Que é Cloud Computing?



A Computação em Nuvem (Cloud Computing) é a entrega de serviços de computação - incluindo servidores, armazenamento, bancos de dados, rede, software, análise e inteligência - pela Internet (“a nuvem”) para oferecer recursos flexíveis, inovação e economia de escala.





O Que é Cloud Computing?



Normalmente, pagamos apenas pelos serviços em nuvem que usamos, ajudando a reduzir os custos operacionais, operar a infraestrutura de forma mais eficiente e escalar conforme as necessidades de negócios mudam.

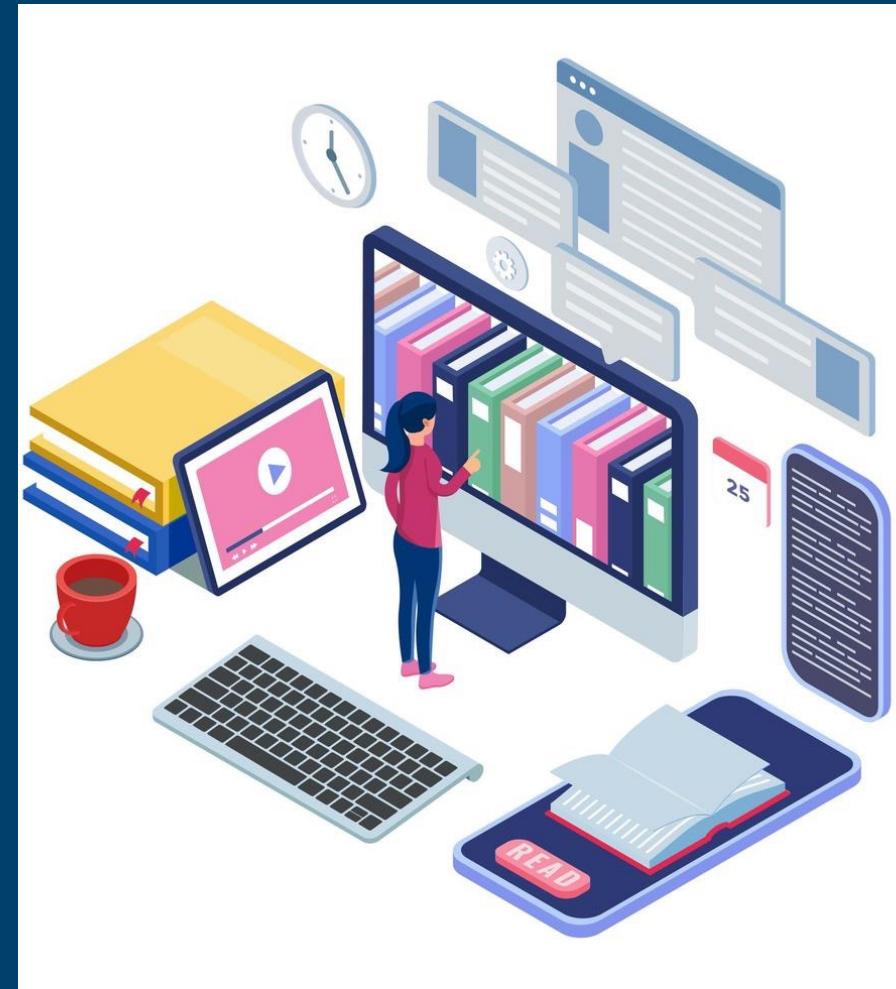


Big Data Fundamentos 3.0

Cloud Computing e Big Data

Data Science Academy

Data Science Academy



Cloud Computing e Big Data



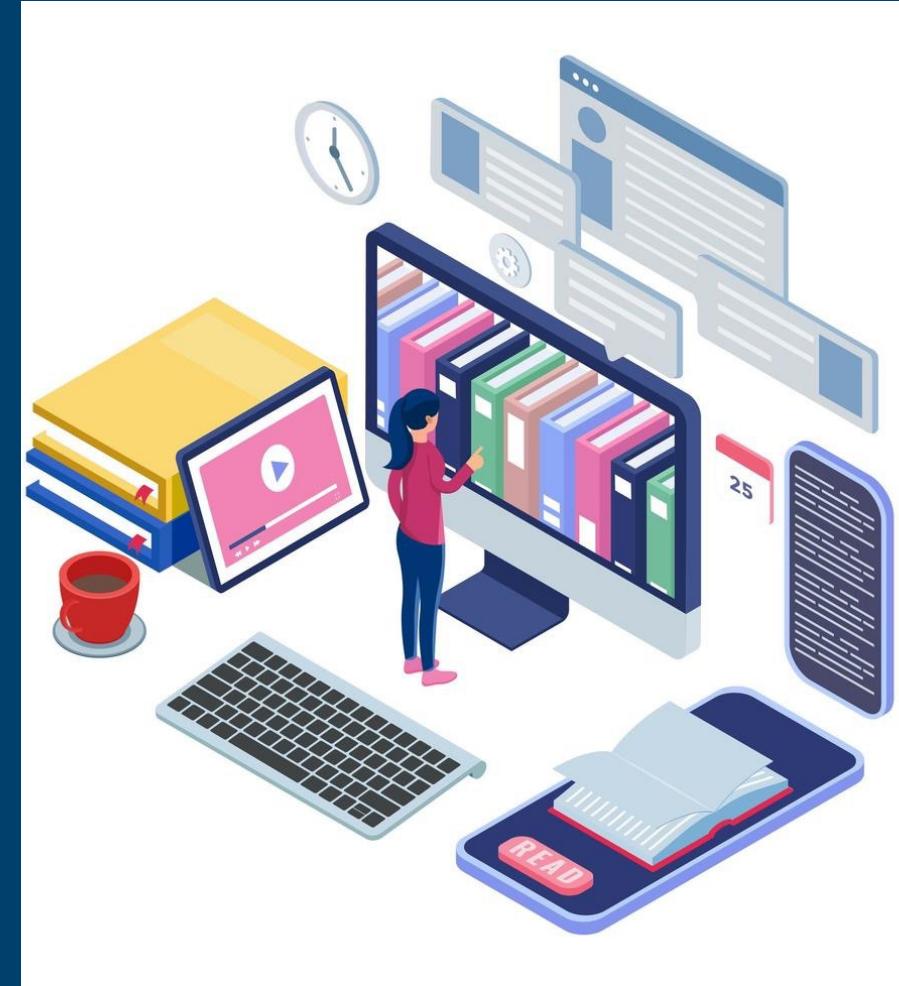
Big Data

Fundamentos 3.0

MLOps e DataOps

Data Science Academy

Data Science Academy



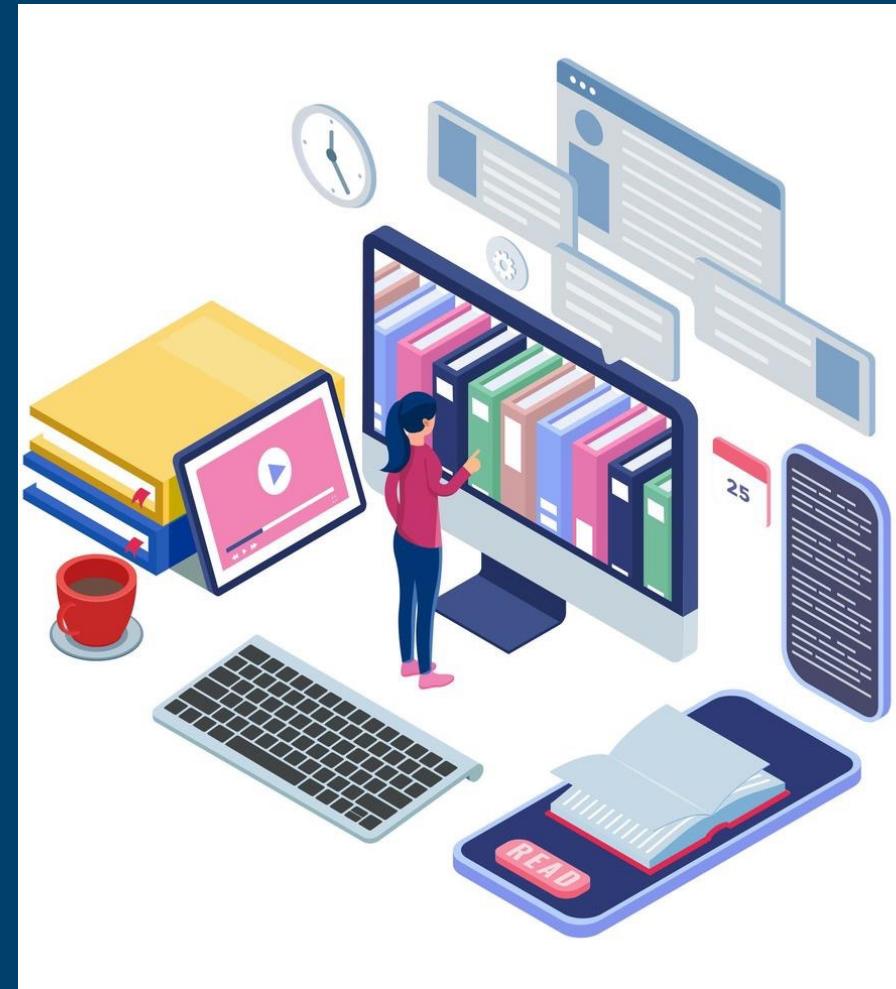
Big Data

Fundamentos 3.0

O Que é Machine Learning?

Data Science Academy

Data Science Academy





O Que é Machine Learning?

O Que é Machine Learning?

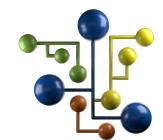
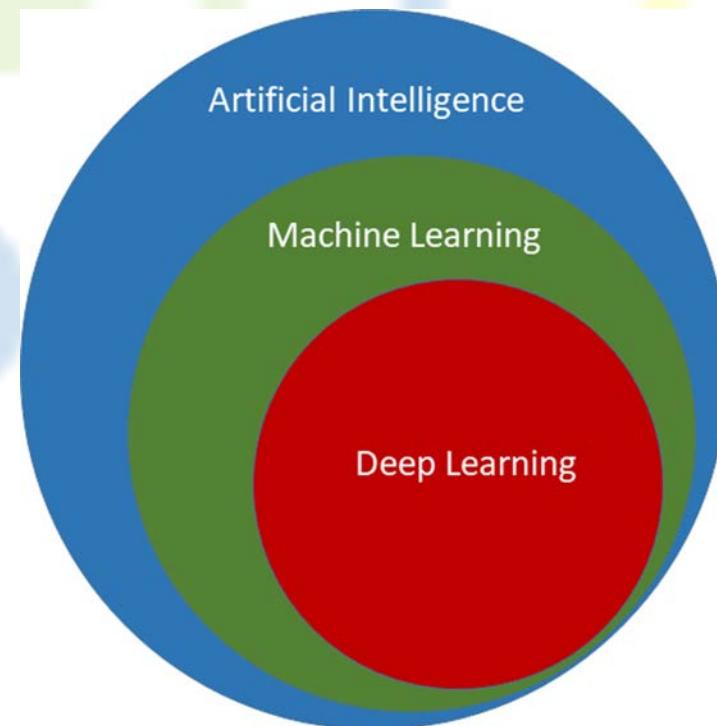
Machine Learning é uma sub-área da Inteligência Artificial (IA) e da Ciência da Computação que se concentra no uso de dados e algoritmos para imitar a forma como os humanos aprendem, melhorando gradativamente sua precisão.





O Que é Machine Learning?

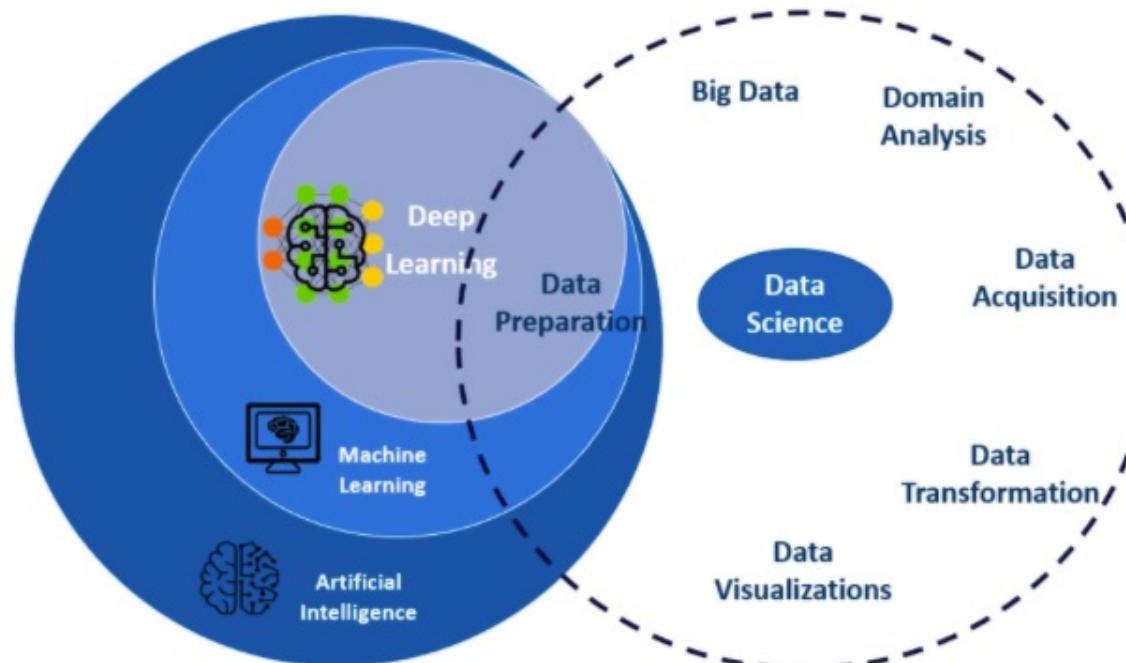
O Que é Machine Learning?





O Que é Machine Learning?

O Que é Machine Learning?





O Que é Machine Learning?

O Que é Machine Learning?

Dados

```
100100011101000000101000110111010110  
10010011110111000000111100110100100  
100001101101111101010011100001101001  
1111110100001101110010101011100001011  
1100111110111111100100001110110110  
01000011010011011000011000100010000  
0101011100110011101100101010111  
001000010101100101000001000010011110  
011101001111110010111010101010111100  
10001000010111000101011101010111000101  
010010000100101011110011100001010000  
010110000010011101010010101110110001  
01101111101011110001010001010001000010000  
011010011011011010001000101111001101  
000101000001100110001100100010010110  
100101010100010011100101010101111101
```

Algoritmo



Modelo

$$f(\mathbf{x})$$

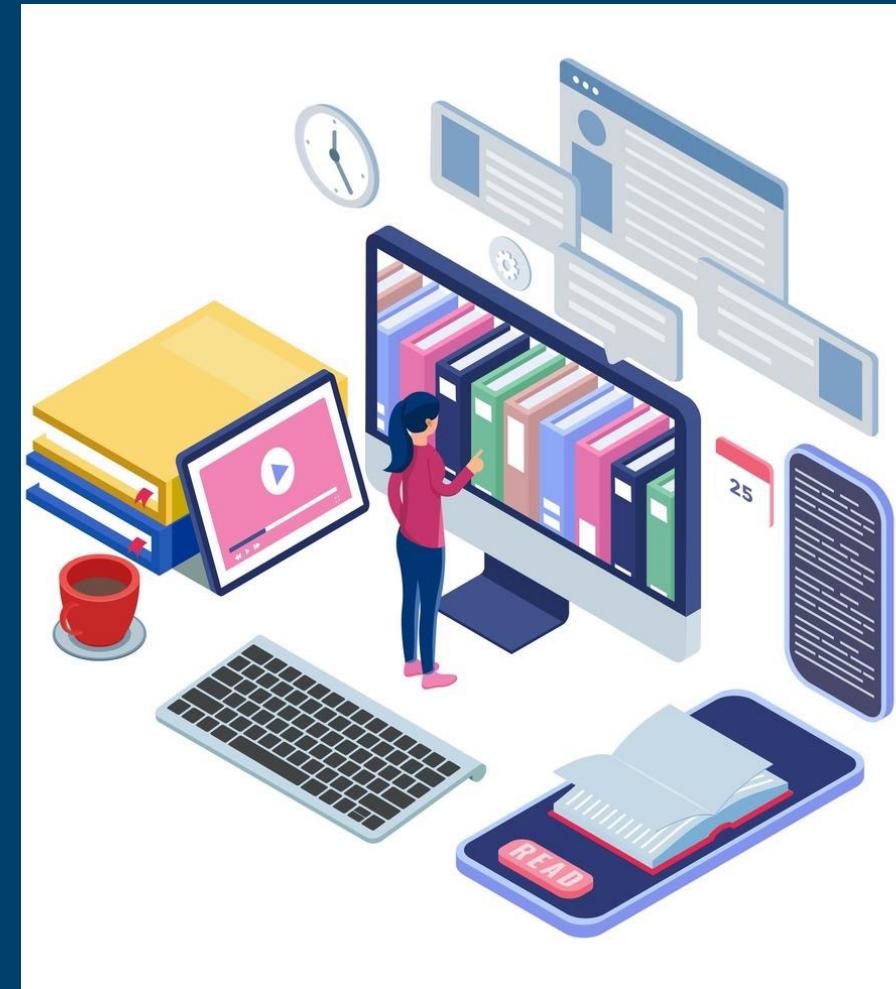


Big Data Fundamentos 3.0

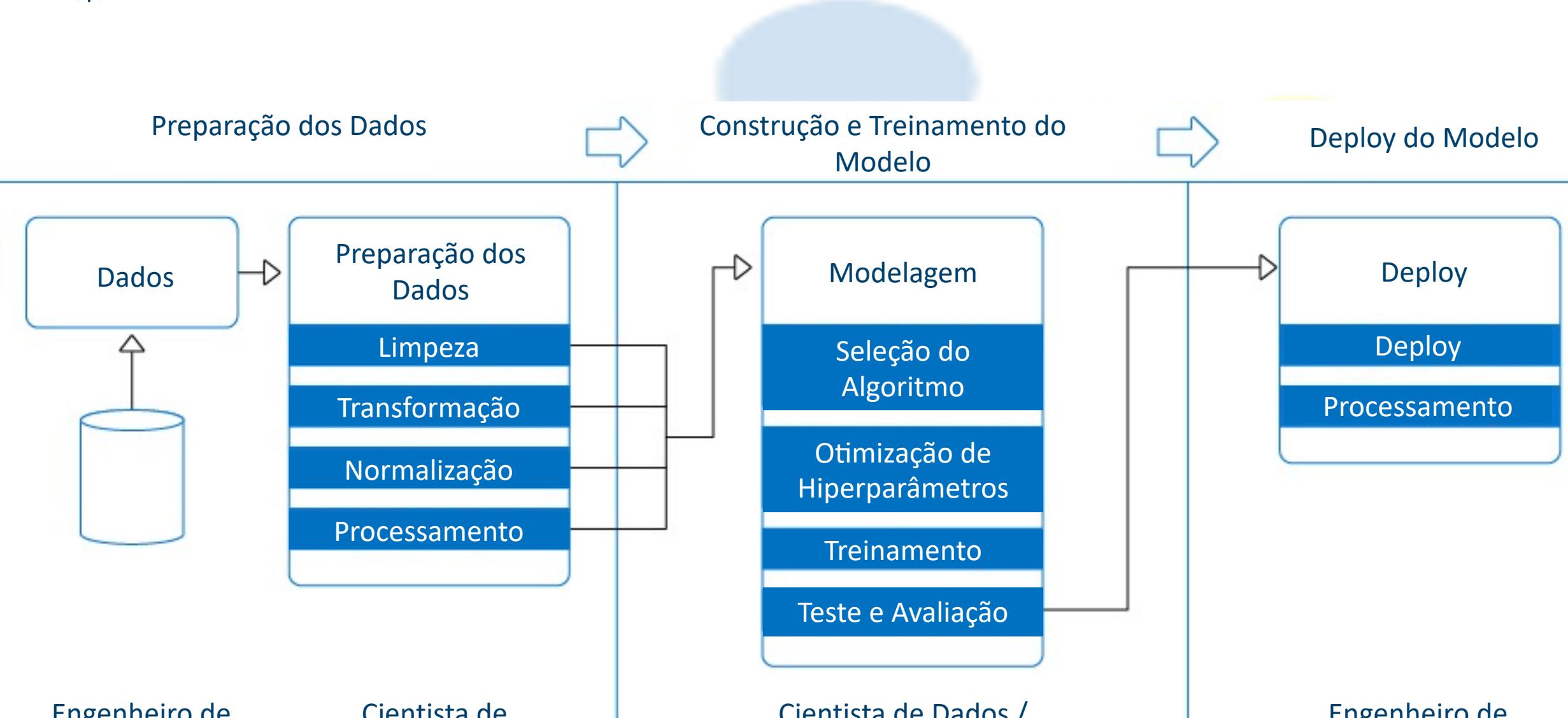
O Pipeline de Machine Learning

Data Science Academy

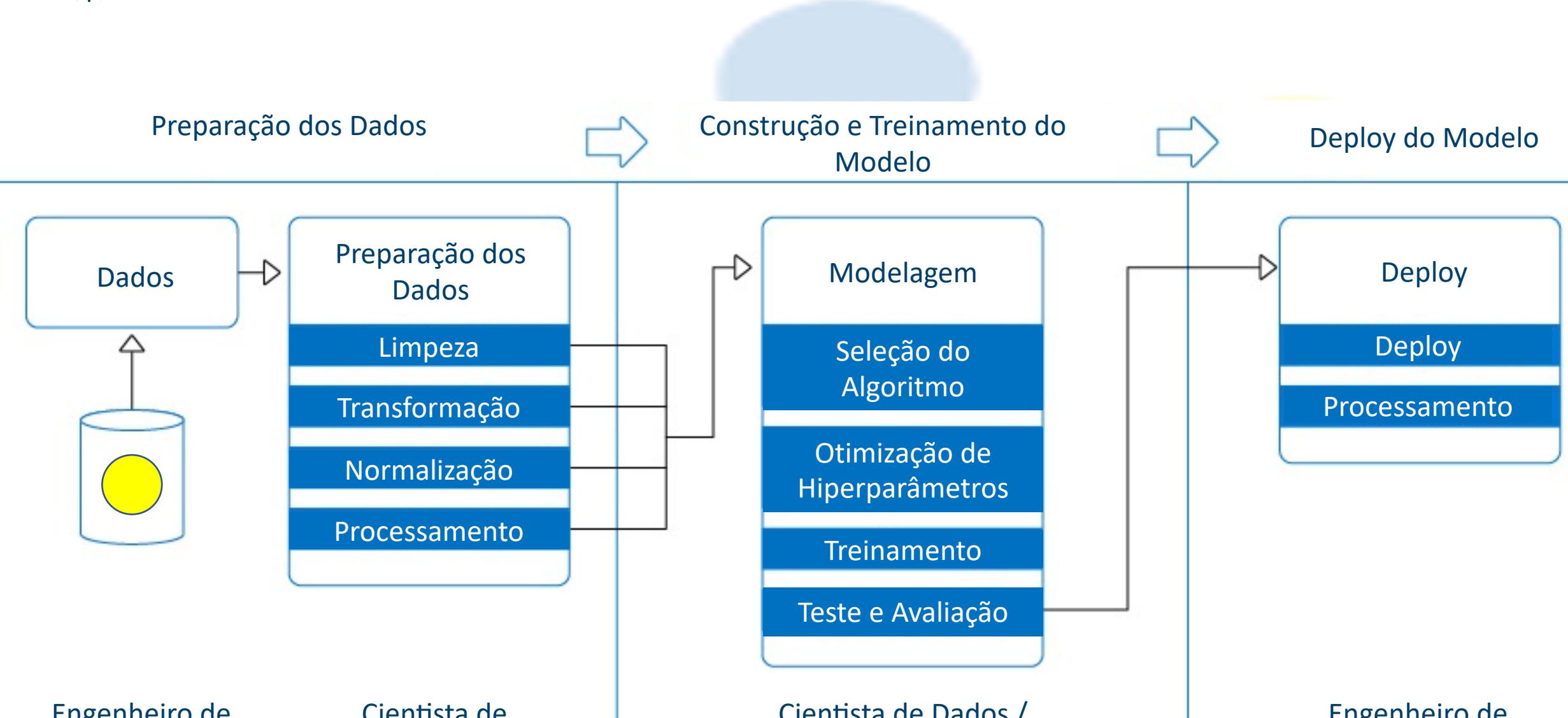
Data Science Academy



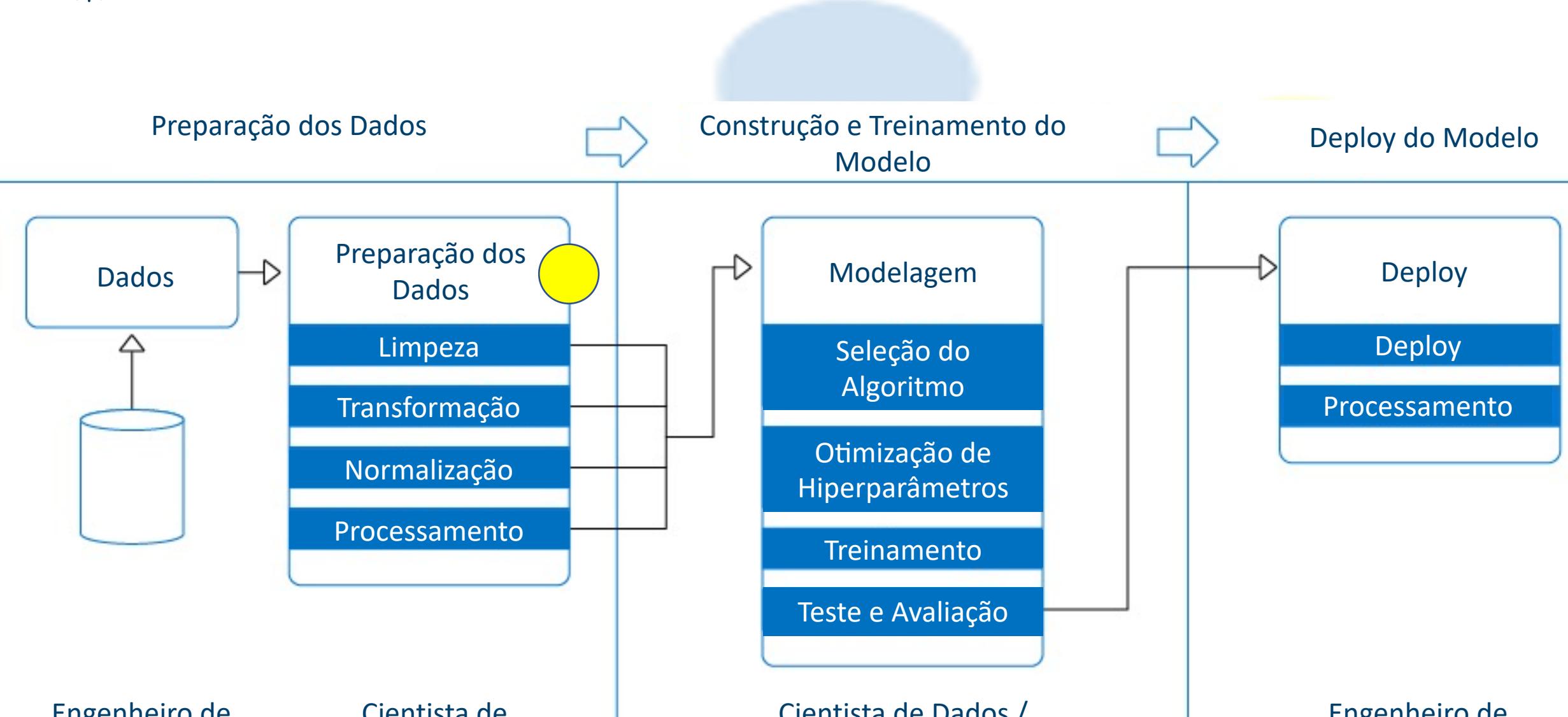
O Pipeline de Machine Learning



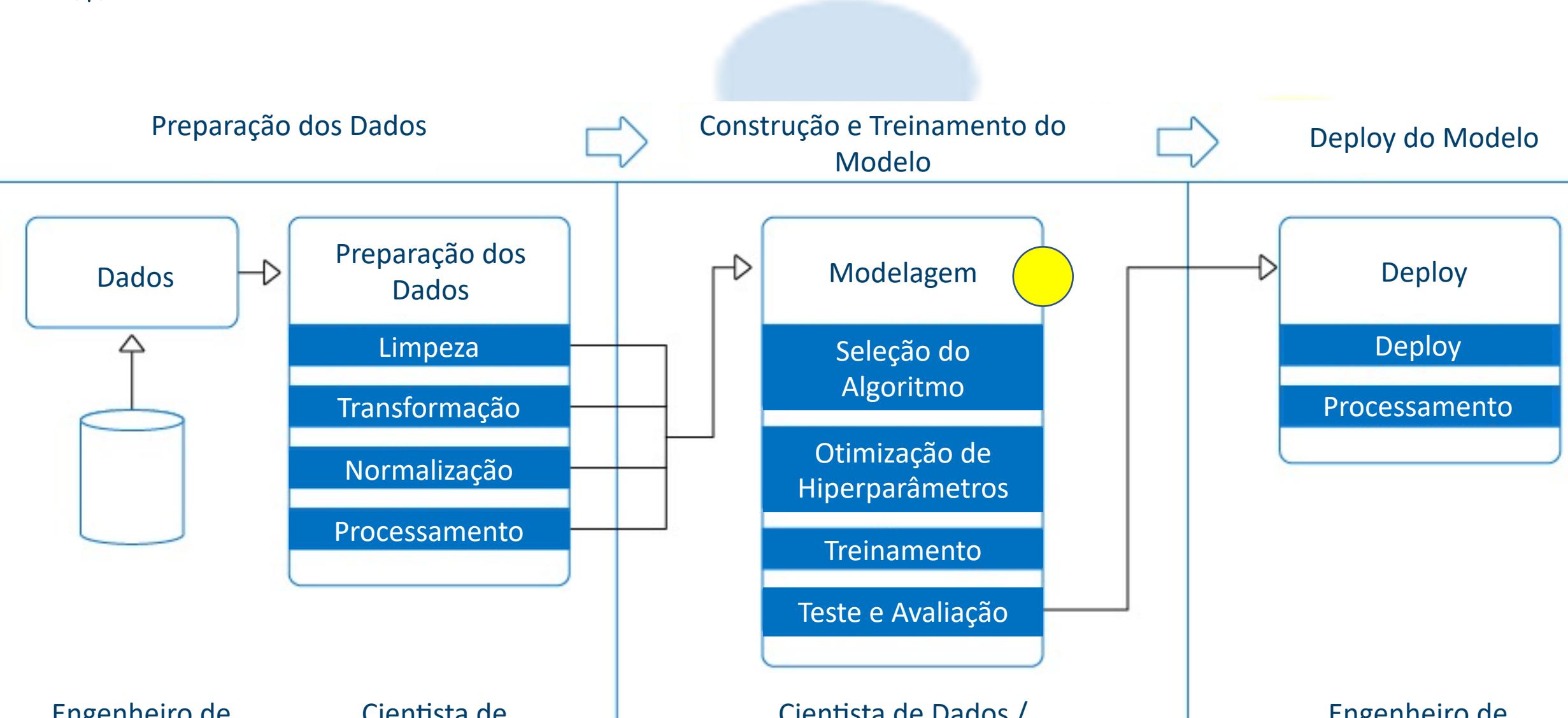
O Pipeline de Machine Learning



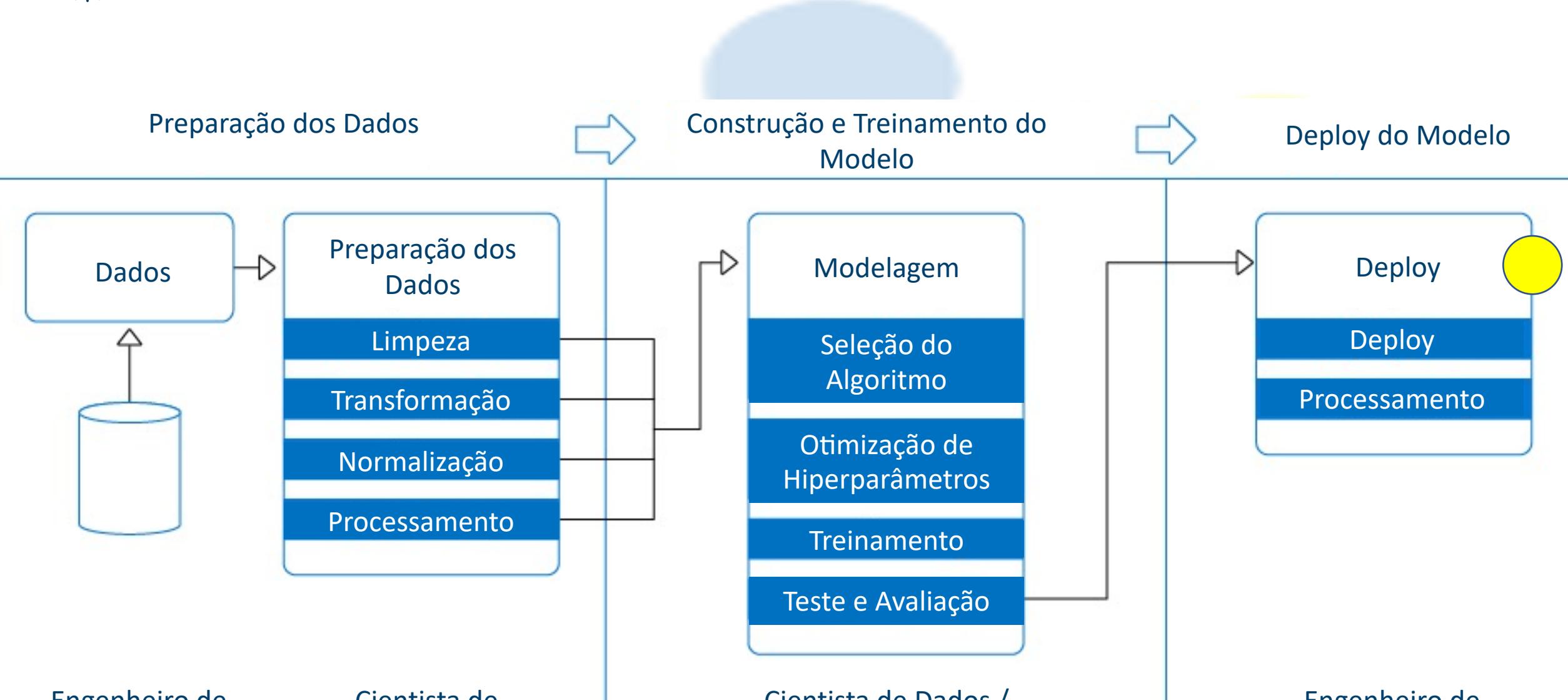
O Pipeline de Machine Learning



O Pipeline de Machine Learning



O Pipeline de Machine Learning



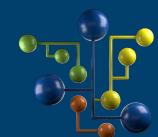
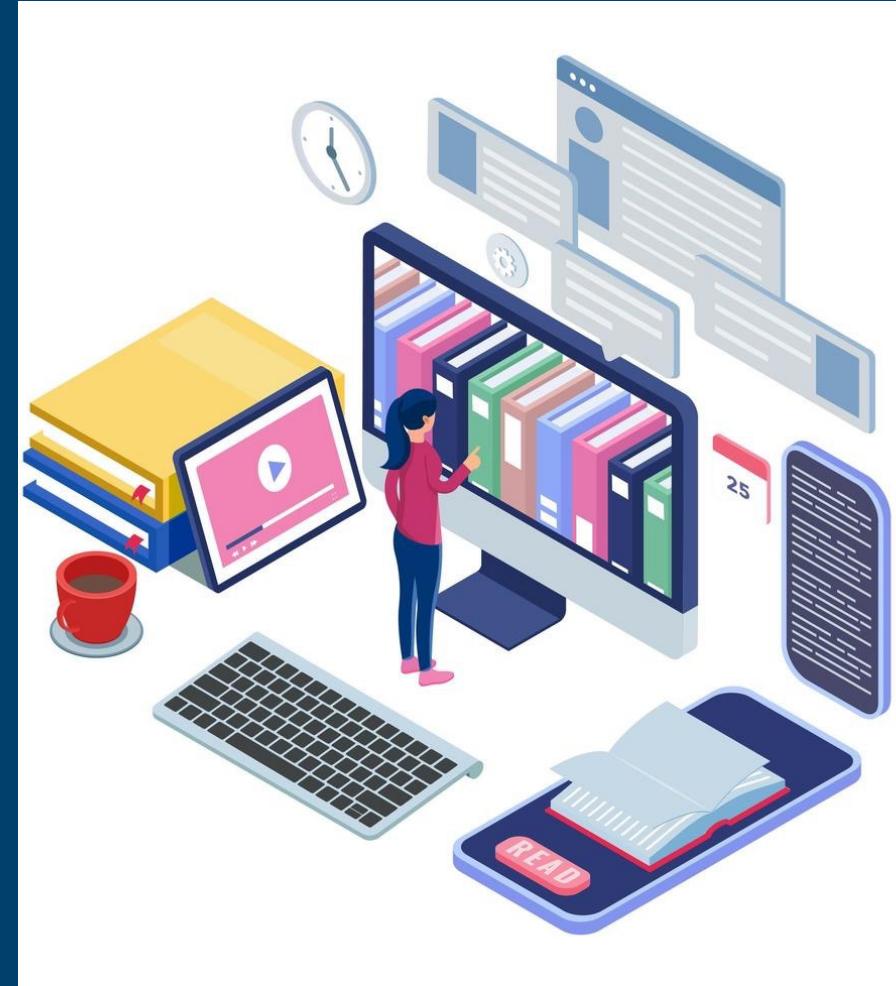
Big Data

Fundamentos 3.0

O Que é Machine Learning Ops?

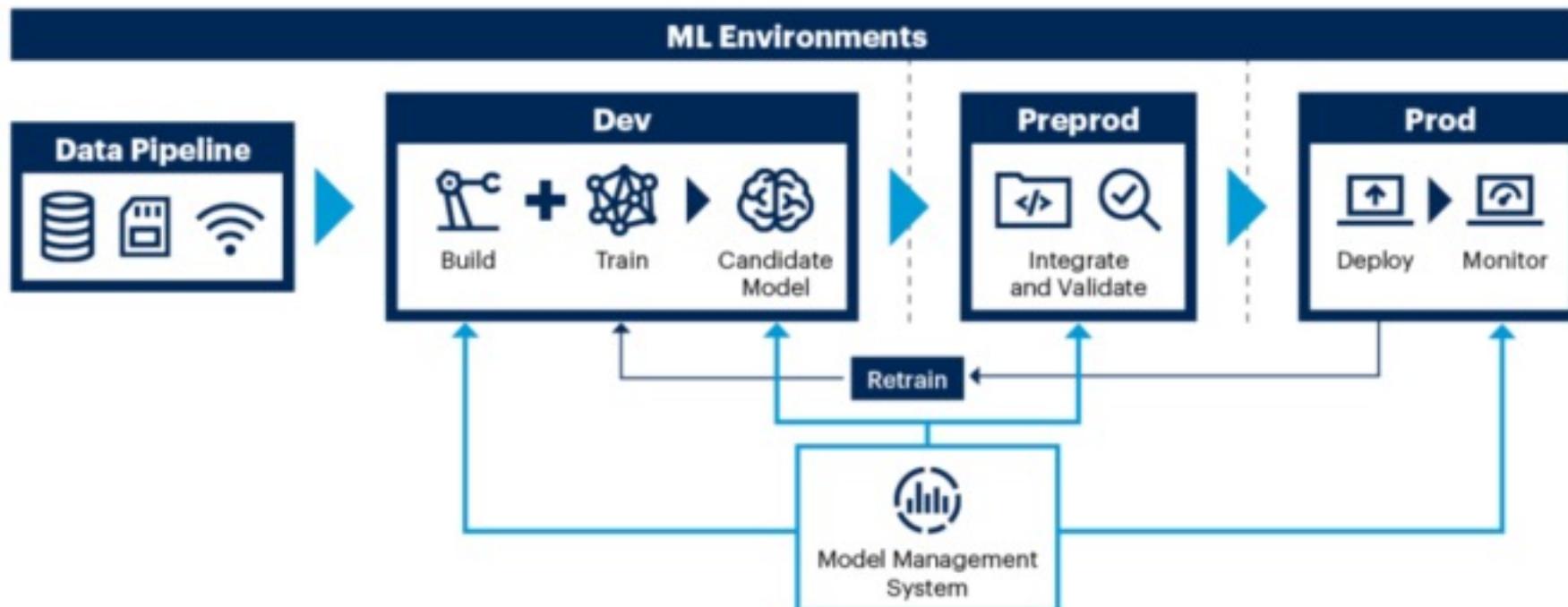
Data Science Academy

Data Science Academy



O Que é Machine Learning Ops?

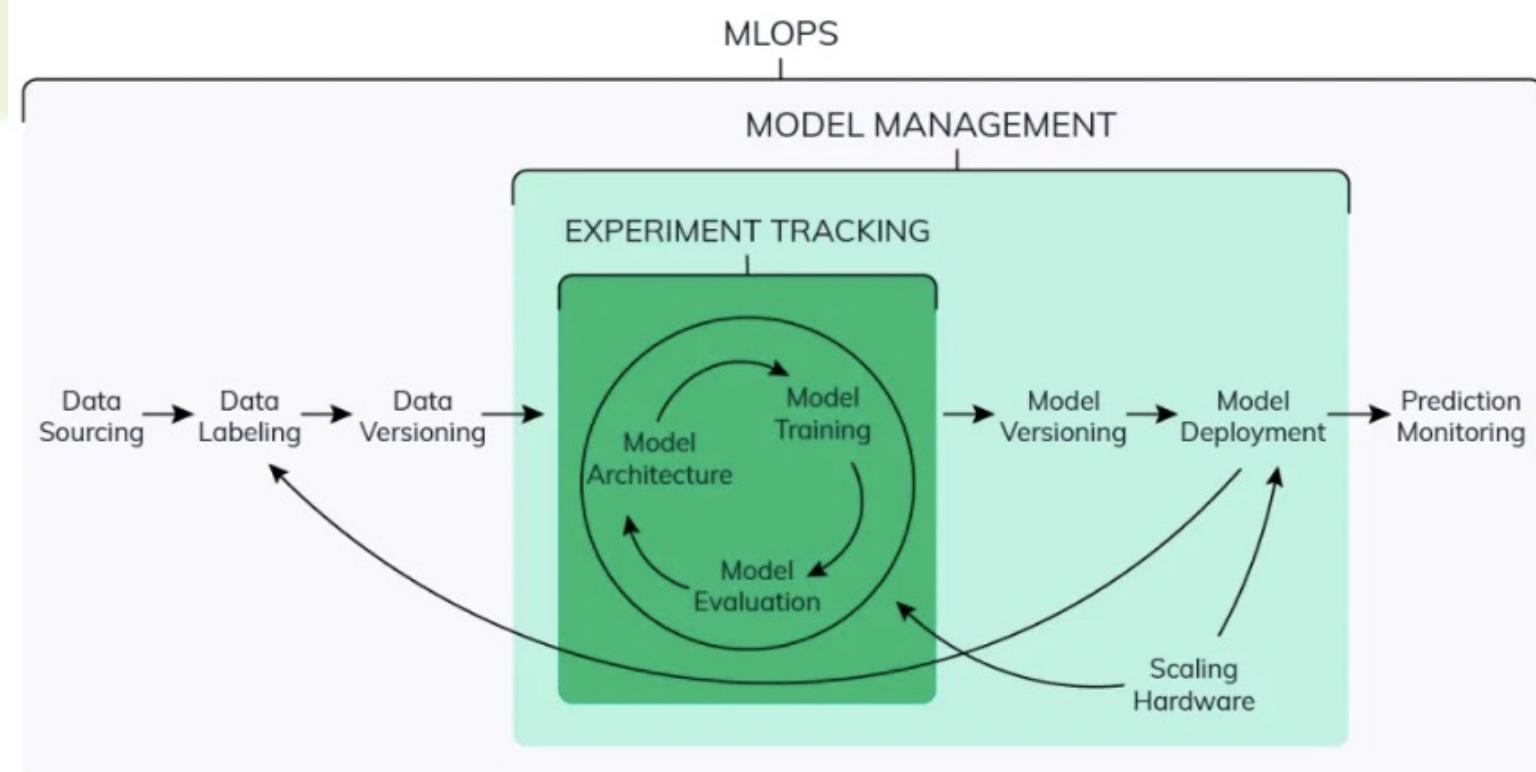
Pipeline de Dados



Source: Gartner



O Que é Machine Learning Ops?





O Que é Machine Learning Ops?

MLOps = ML + DEV + OPS





O Que é Machine Learning Ops?

MLOps é um conjunto de práticas para colaboração e comunicação entre Cientistas de Dados e profissionais de operações.

MLOps é normalmente tarefa do Engenheiro de Machine Learning.

A aplicação dessas práticas aumenta a qualidade, simplifica o processo de gerenciamento e automatiza a implantação de modelos de aprendizado de máquina em ambientes de produção em grande escala. É mais fácil alinhar os modelos às necessidades de negócios, bem como aos requisitos regulamentares.

MLOps visa unificar o desenvolvimento de sistemas de ML (dev) e a implantação de sistemas de ML (ops) para padronizar e agilizar a entrega contínua de modelos de alto desempenho em produção.



Big Data

Fundamentos 3.0

DevOps, MLOps, AIOps, DataOps

Data Science Academy

Data Science Academy





DevOps, MLOps, AIOps, DataOps

DevOps é uma abordagem para desenvolvimento de software que acelera o ciclo de vida de construção usando automação. O DevOps se concentra na implantação contínua de software, aproveitando os recursos de TI sob demanda e automatizando a integração, o teste e a implantação de código. Essa fusão de desenvolvimento de software (“dev”) e operações de TI (“ops”) reduz o tempo de implantação, diminui o tempo de lançamento no mercado, minimiza defeitos e diminui o tempo necessário para resolver problemas.





DevOps, MLOps, AIOps, DataOps

Usando DevOps, empresas conseguiram reduzir o tempo do ciclo de lançamento de software de meses para literalmente segundos. Essa descoberta permitiu o crescimento e liderança em mercados emergentes e em ritmo acelerado. Empresas como Google, Amazon e muitas outras agora lançam software muitas vezes por dia. Ao melhorar a qualidade e o tempo de ciclo dos lançamentos de código, o DevOps merece muito crédito pelo sucesso dessas empresas.



DevOps, MLOps, AIOps, DataOps

Várias empresas se especializaram em DevOps ao longo do tempo e diversas novas ferramentas surgiram.

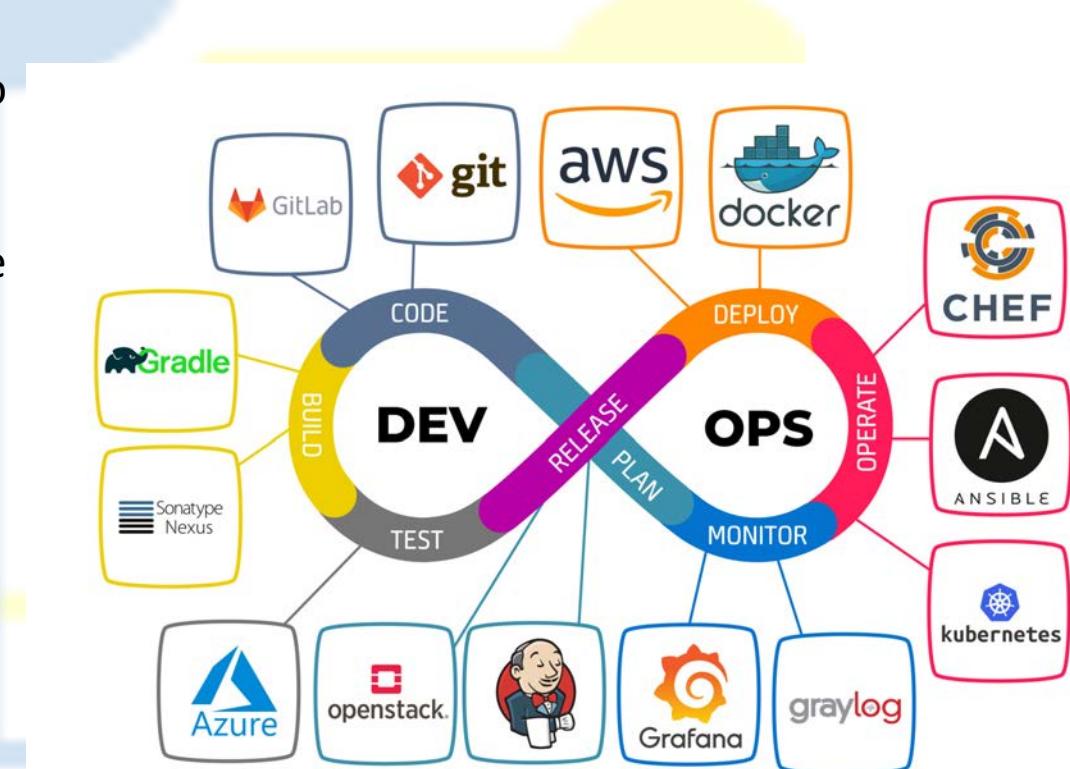
Então, por que não levar o mesmo conceito para a Ciência de Dados?

E assim nasceram:

MLOps – Operação do fluxo de trabalho em Machine Learning.

AIOps – Operação do fluxo de trabalho em IA.

DataOps – Conceito mais recente que abrange toda a operação de dados de uma empresa.



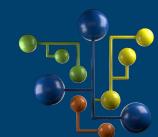
Big Data

Fundamentos 3.0

O Que é DataOps?

Data Science Academy

Data Science Academy

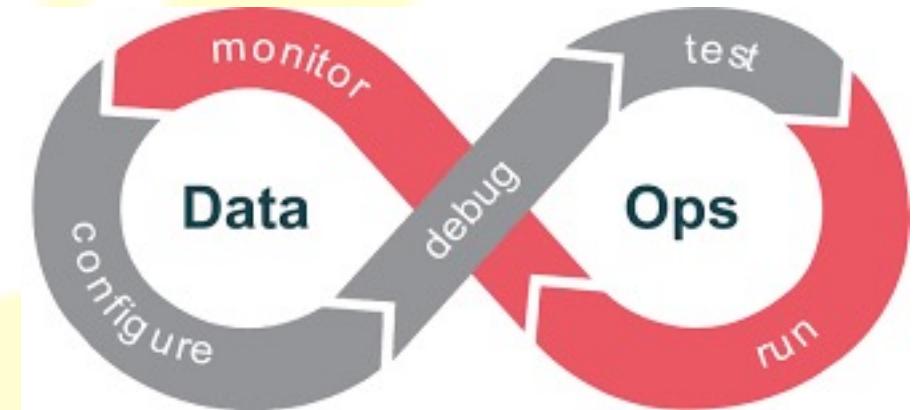


O Que é DataOps?

DataOps (Operações de Dados) é uma metodologia ágil e orientada a processos para desenvolver e entregar análises.

DataOps fornece as ferramentas, processos e estruturas organizacionais para apoiar a empresa focada em dados.

DataOps é a capacidade de habilitar soluções, desenvolver produtos de dados e ativar dados para valor comercial em todas as camadas de tecnologia, da infraestrutura à experiência do usuário final.



O Que é DataOps?

O objetivo do DataOps é agilizar o design, o desenvolvimento e a manutenção de aplicativos com base em dados e análise de dados. Busca melhorar a forma como os dados são gerenciados e os produtos são criados e coordenar essas melhorias com os objetivos do negócio.

As equipes de DataOps também buscam orquestrar dados, ferramentas, código e ambientes do início ao fim, com o objetivo de fornecer resultados reproduzíveis.

As equipes de DataOps tendem a ver os pipelines analíticos como análogos às linhas de produção de uma fábrica, sendo que aqui a matéria-prima é o Big Data.





O Que é DataOps?



Operações de Dados



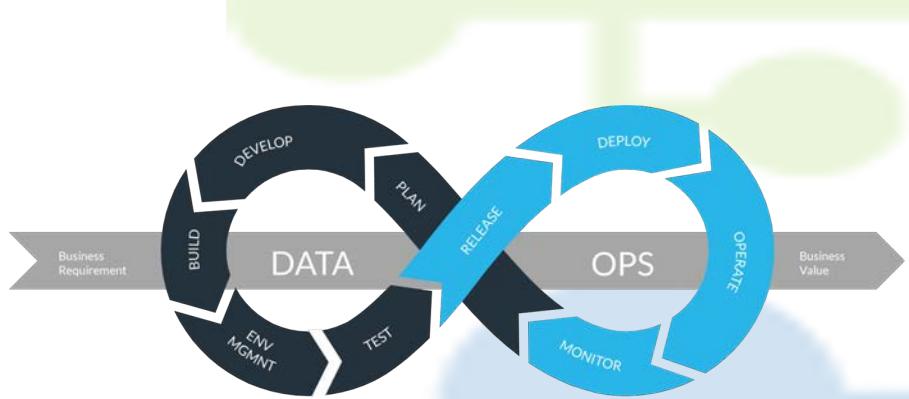
Produto Final



Data Science Academy



O Que é DataOps?



Operações de Dados



Produto Final



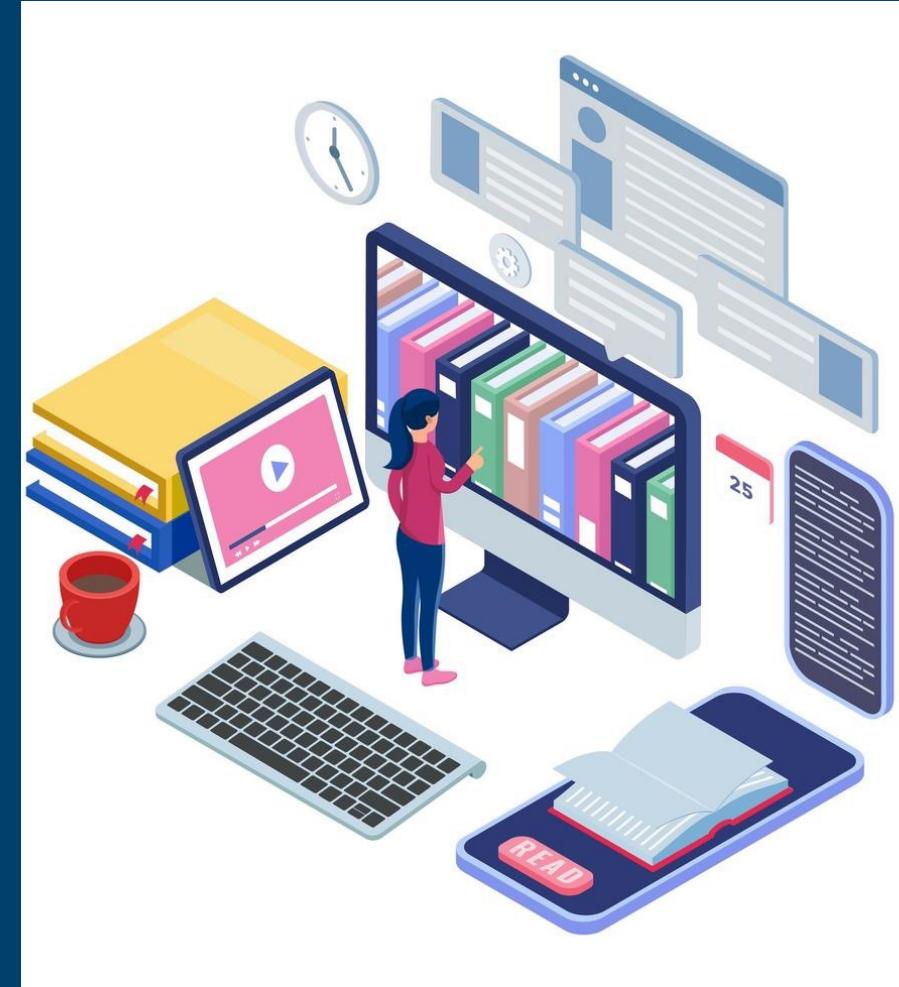
Big Data

Fundamentos 3.0

Big Data x Small Data

Data Science Academy

Data Science Academy





Big Data x Small Data

Big Data

Grandes volumes de dados,
com muita variedade e
gerados em alta velocidade.

Small Data

Dados que estão disponíveis
em quantidade mínima
suficiente para compreensão
humana.



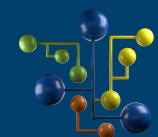
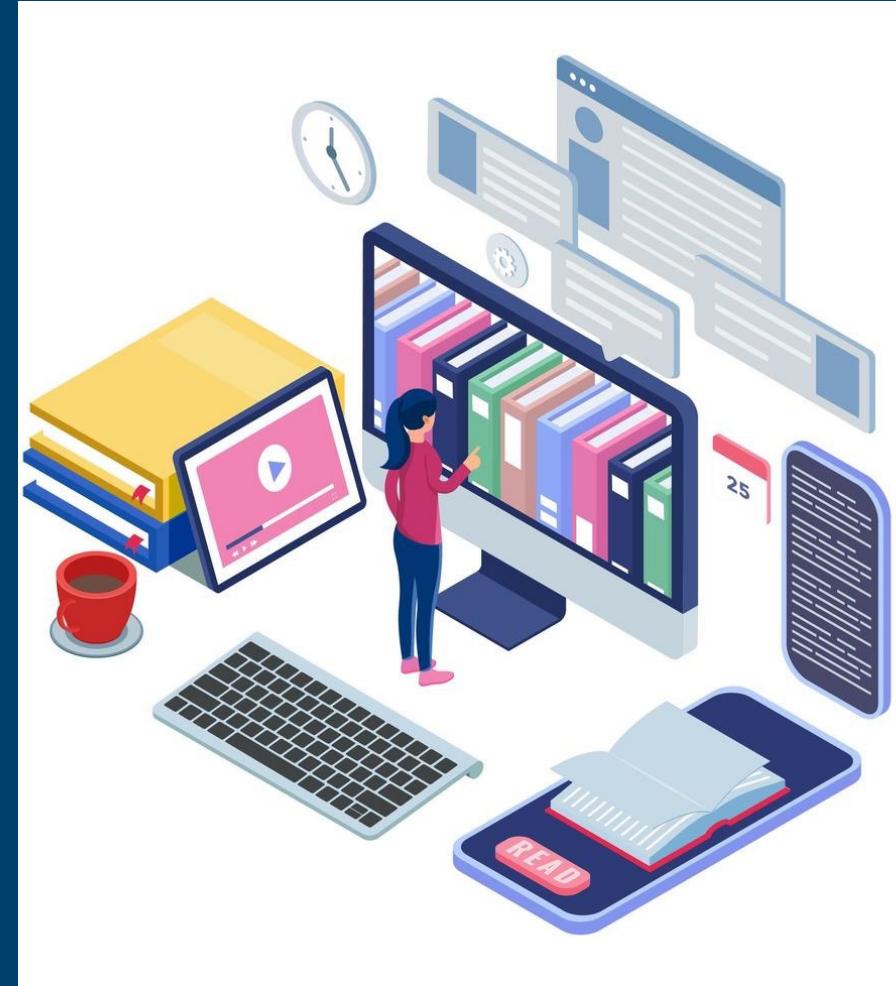
Big Data

Fundamentos 3.0

Dados Como Serviço

Data Science Academy

Data Science Academy



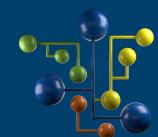
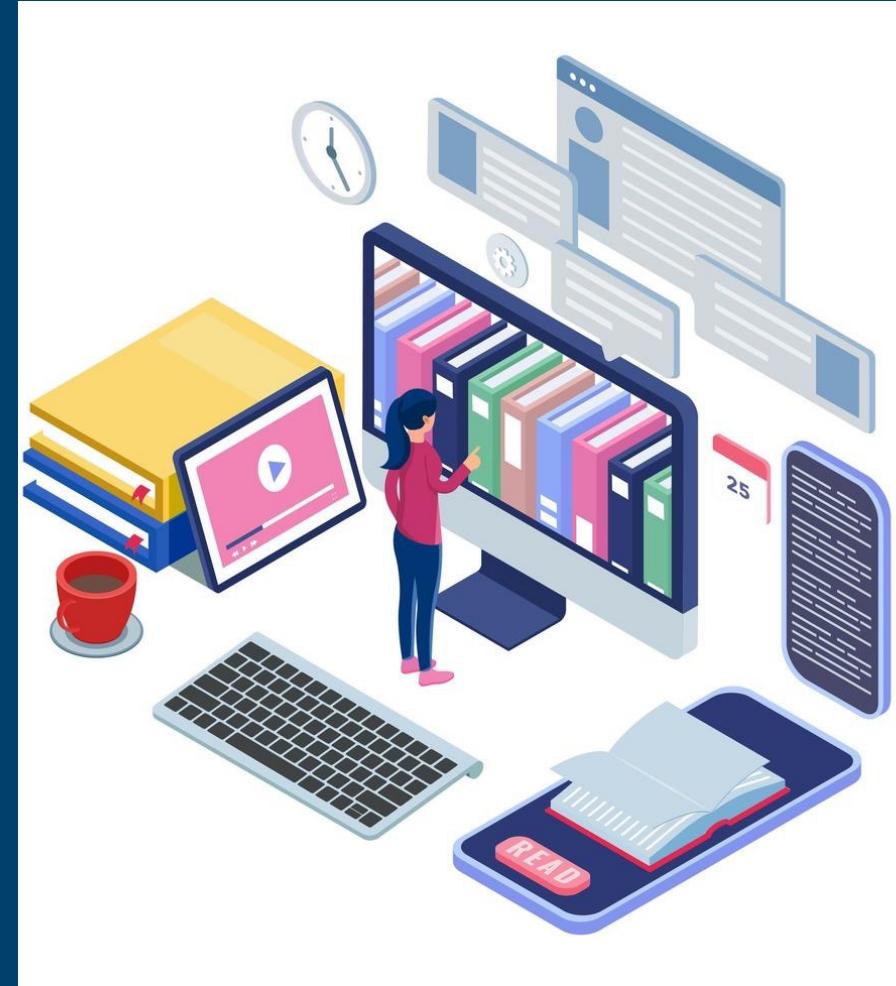
Big Data

Fundamentos 3.0

Data as a Service (DaaS)

Data Science Academy

Data Science Academy

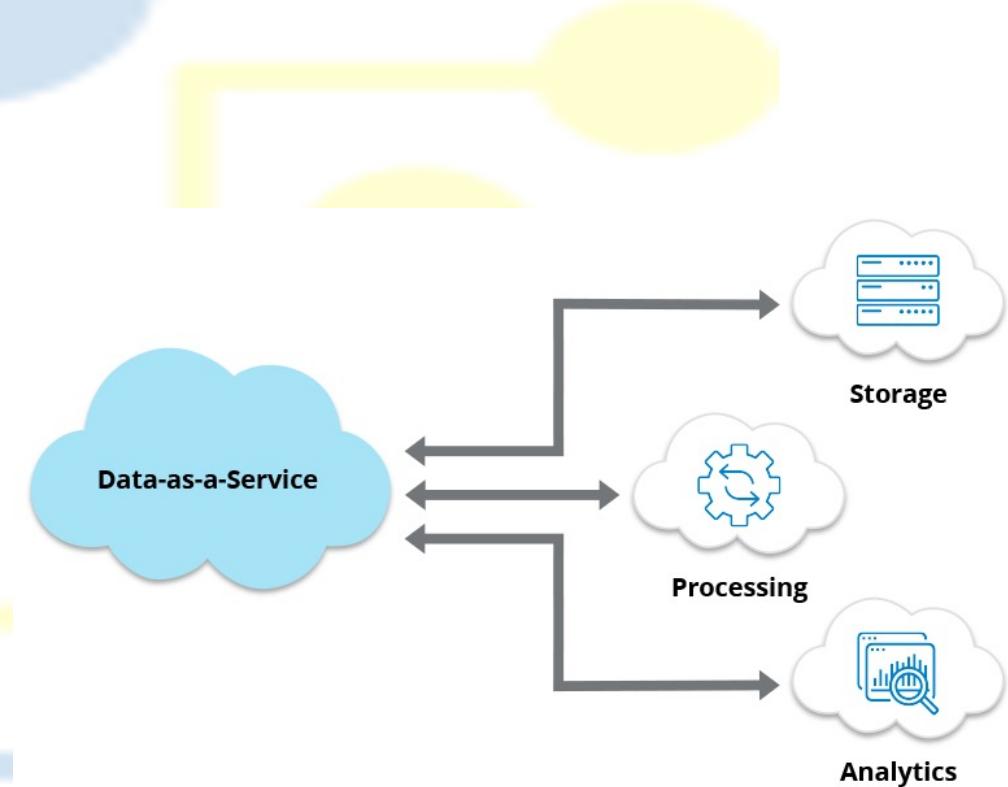


Data as a Service (DaaS)

Data as a Service (DaaS) é uma estratégia de gerenciamento de dados que visa alavancar os dados como um ativo de negócios para maior agilidade no processo de análise.

Faz parte das ofertas “as a service” que se tornaram cada vez mais populares desde a expansão da Internet nos anos 1990, que começou com a introdução do Software as a Service (SaaS).

Semelhante a outros modelos “como serviço”, o DaaS fornece uma maneira de gerenciar as grandes quantidades de dados que as organizações geram todos os dias e fornecer essas informações valiosas em toda a empresa para a tomada de decisões baseada em dados.



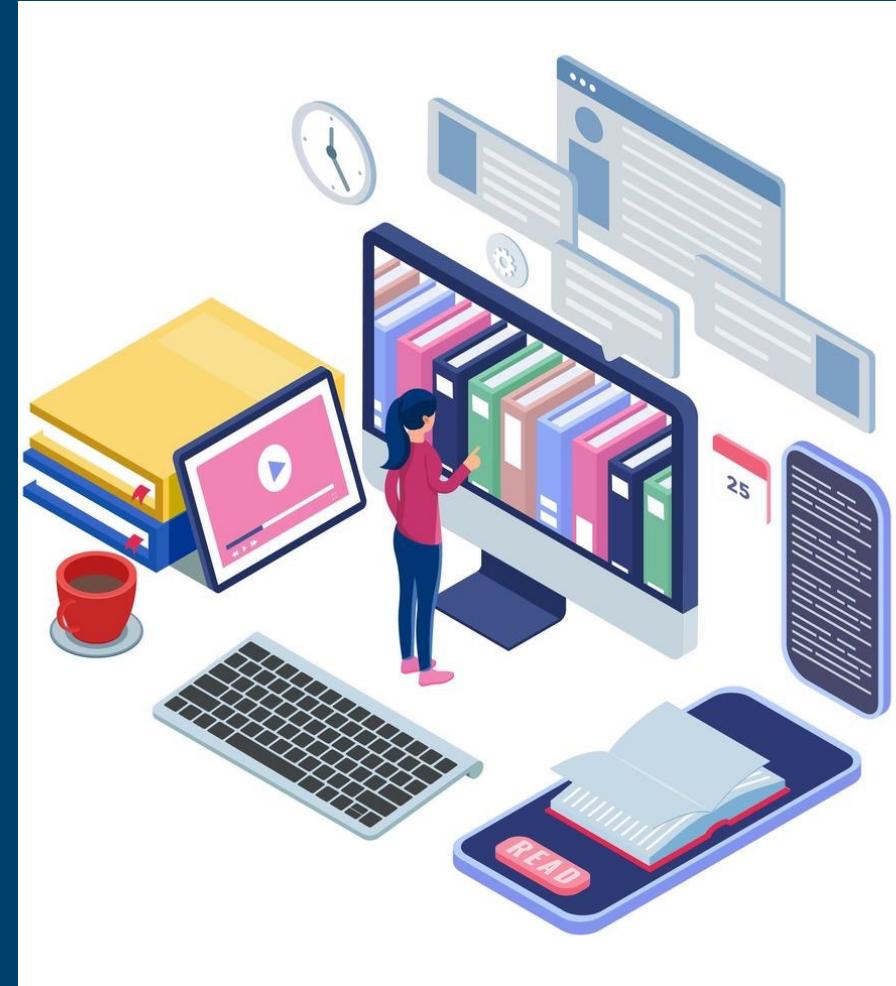
Big Data

Fundamentos 3.0

Arquitetura DaaS

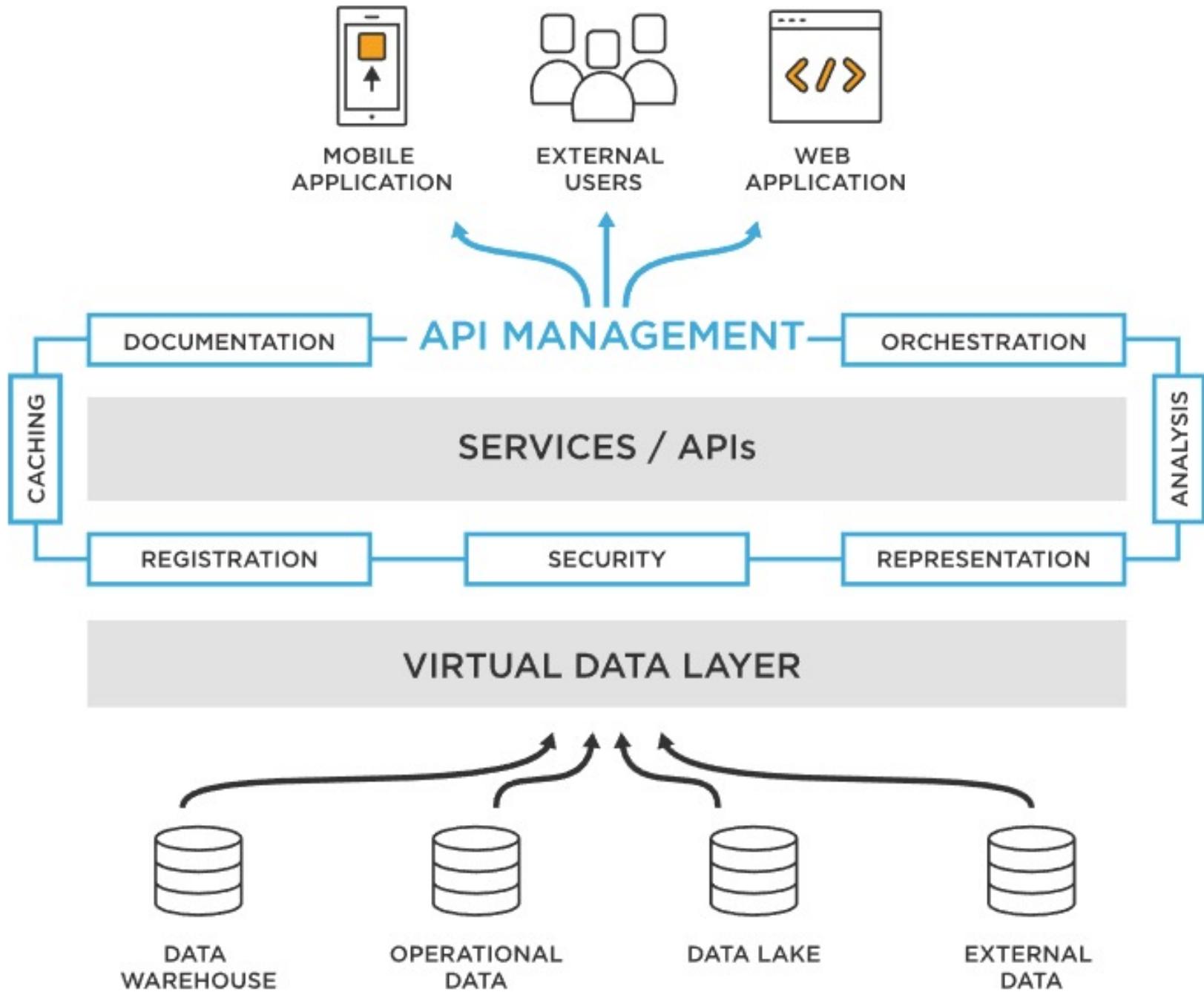
Data Science Academy

Data Science Academy





Arquitetura DaaS

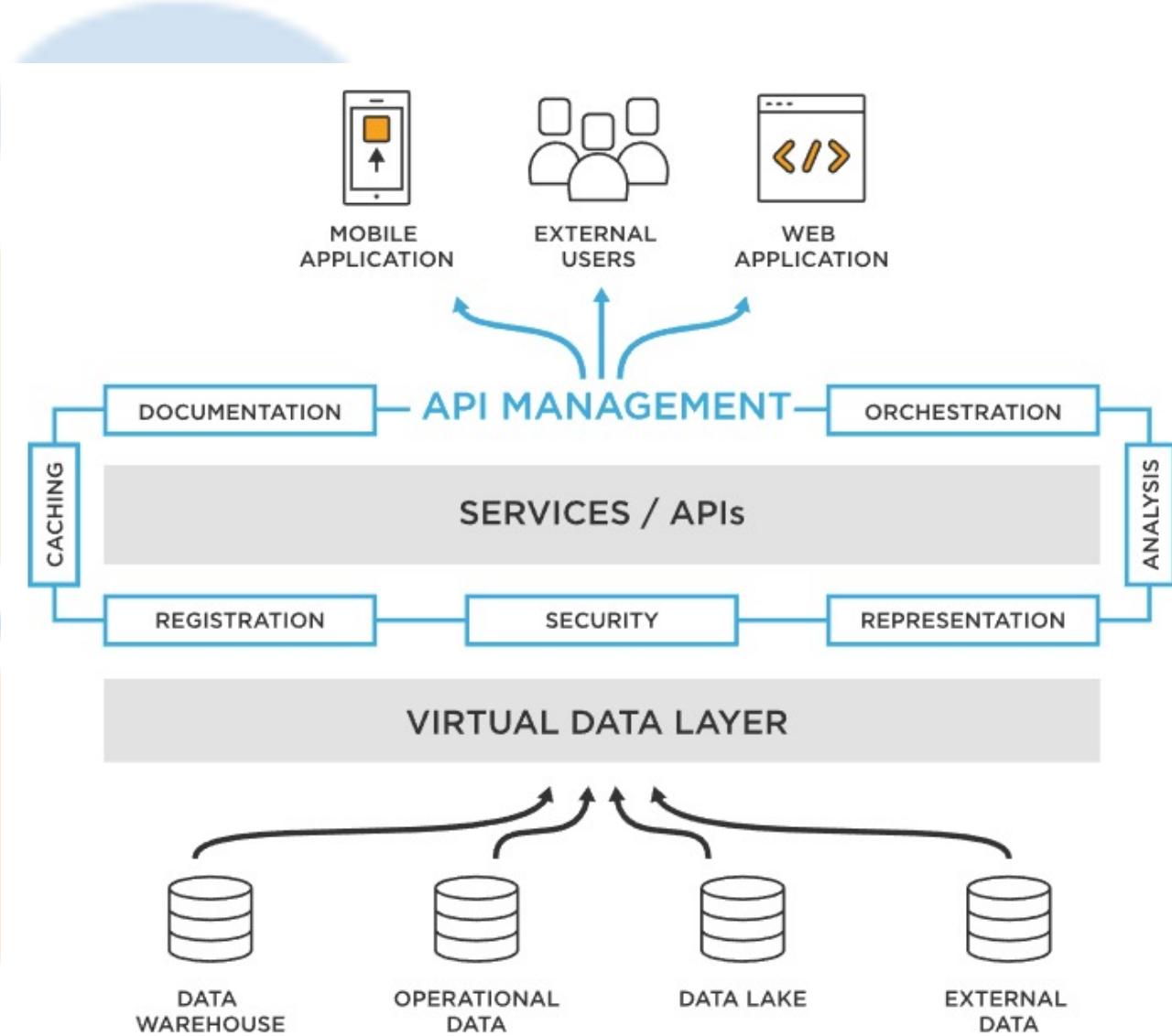


Arquitetura DaaS

A Arquitetura DaaS (Data as a Service) se concentra no provisionamento de dados de uma variedade de fontes sob demanda por meio do uso de APIs.

Projetado para simplificar o acesso aos dados, o DaaS oferece conjuntos de dados já tratados ou fluxos de dados para serem consumidos em uma variedade de formatos, geralmente unificados usando virtualização de dados.

Na verdade, uma Arquitetura DaaS pode incluir uma variedade de tecnologias de gerenciamento de dados, incluindo virtualização de dados, serviços de dados, análise de autoatendimento (Self-Service Analytics) e catalogação de dados.



Big Data Fundamentos 3.0

Principais Benefícios de DaaS

Data Science Academy

Data Science Academy





Principais Benefícios de Daas



Big Data

Fundamentos 3.0

ETL - Extração, Transformação e Carga de Dados

Data Science Academy

Data Science Academy



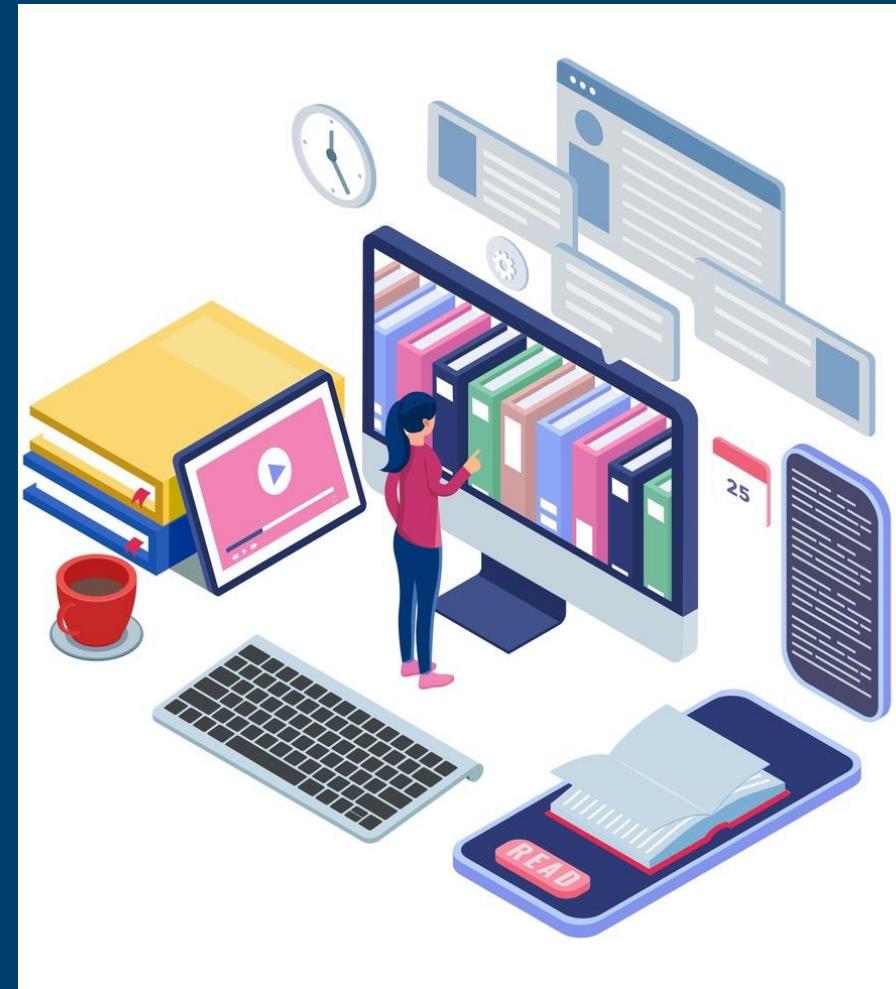
Big Data

Fundamentos 3.0

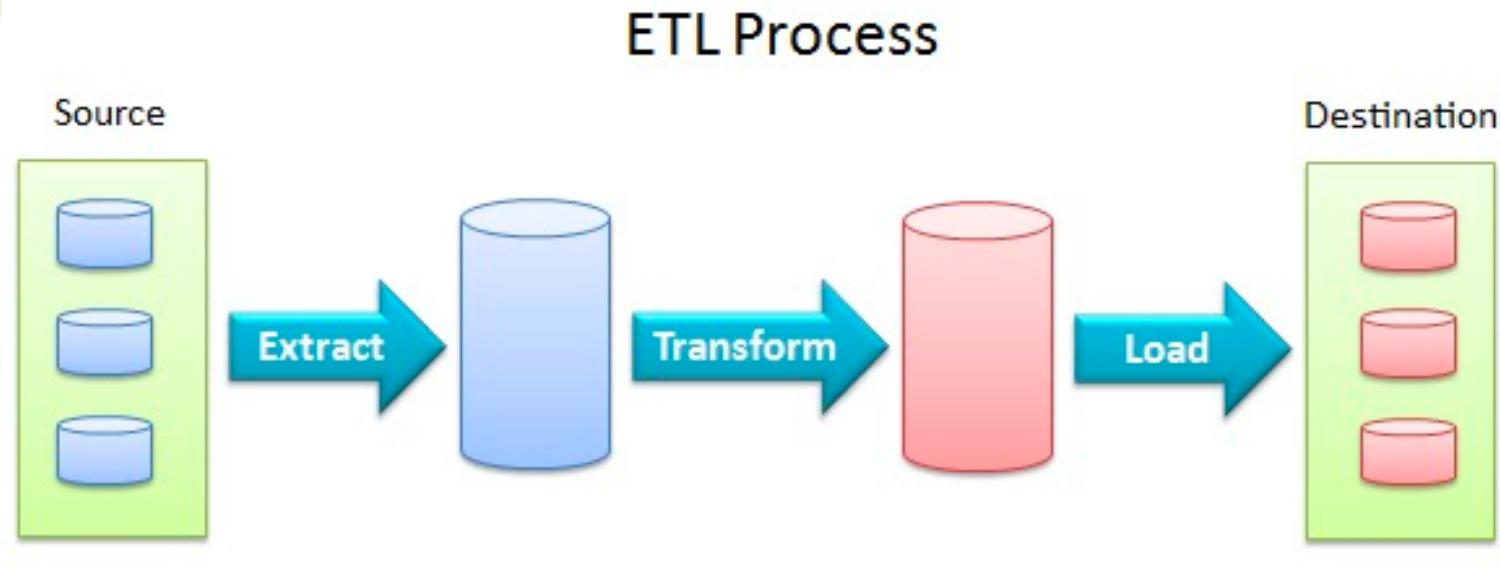
Definindo ETL

Data Science Academy

Data Science Academy

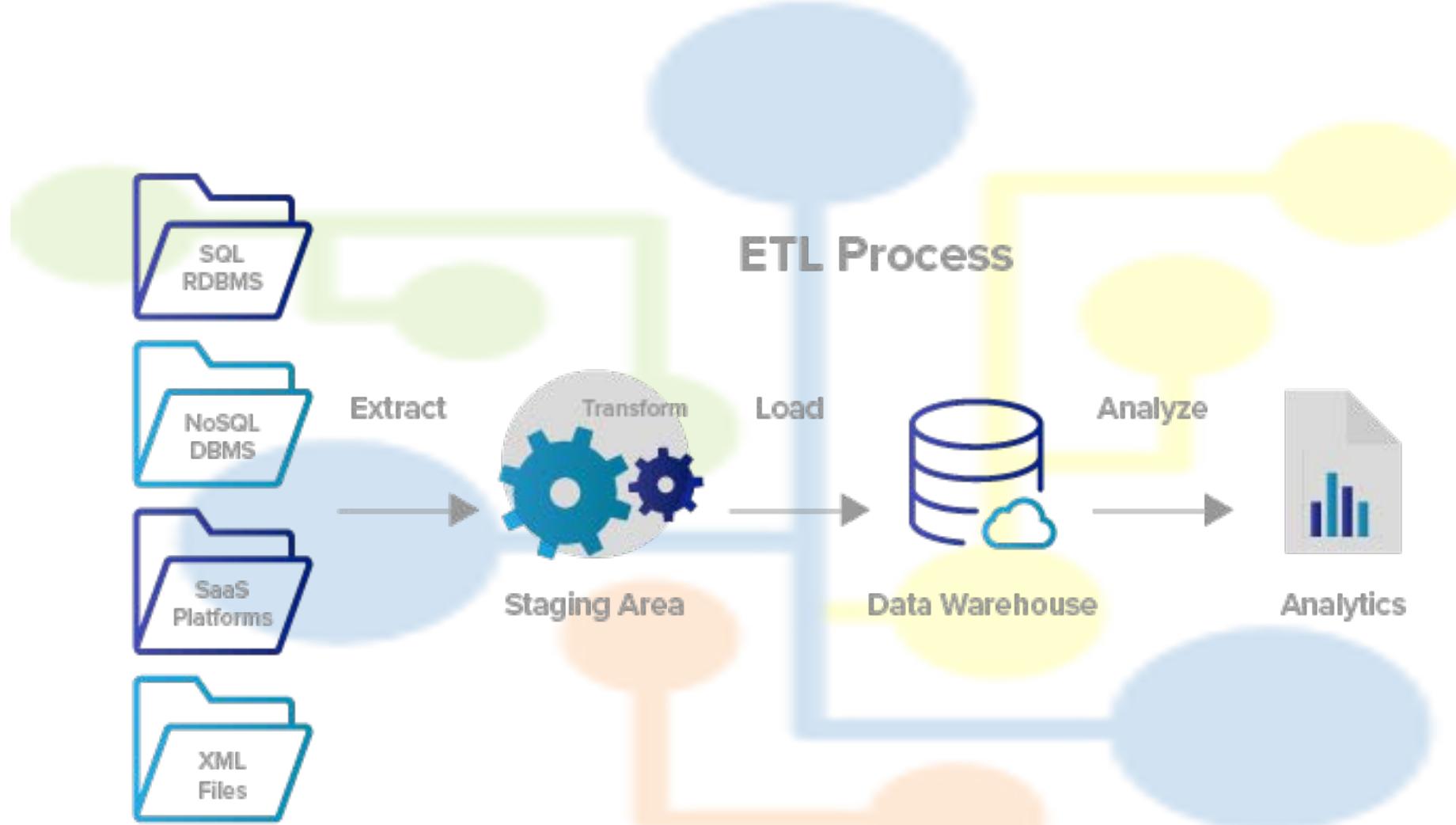


Definindo ETL





Definindo ETL



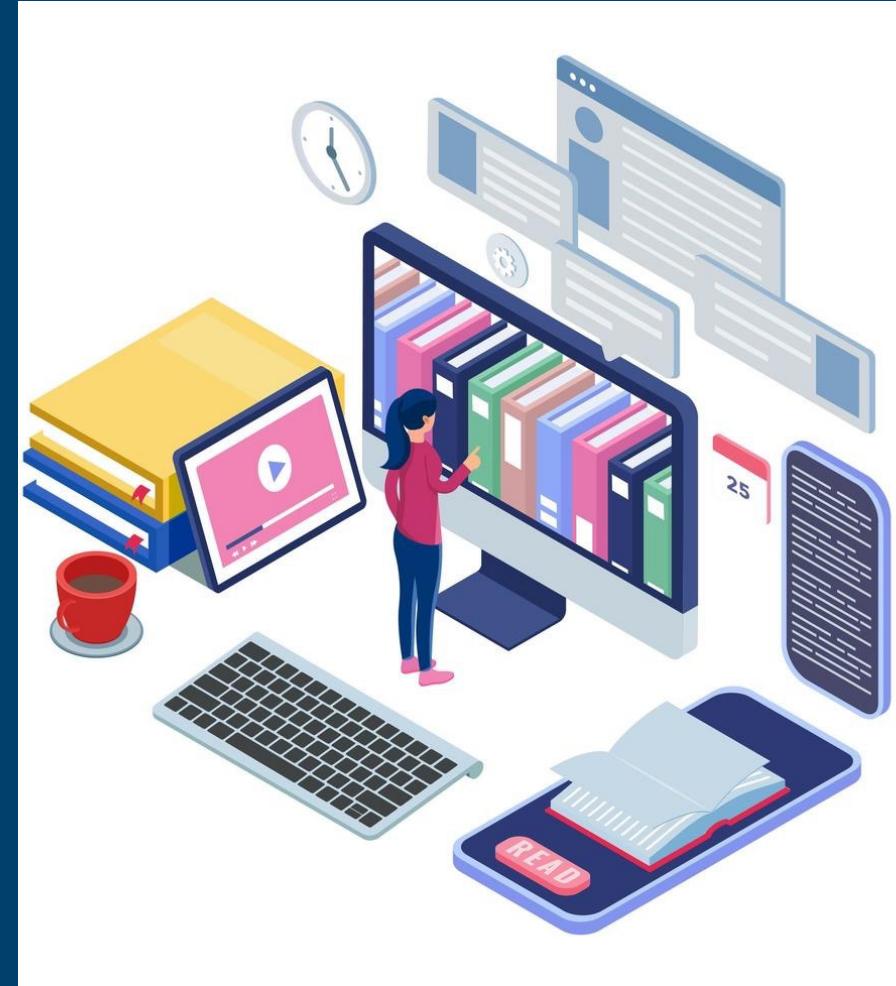
Big Data

Fundamentos 3.0

ETL x ELT

Data Science Academy

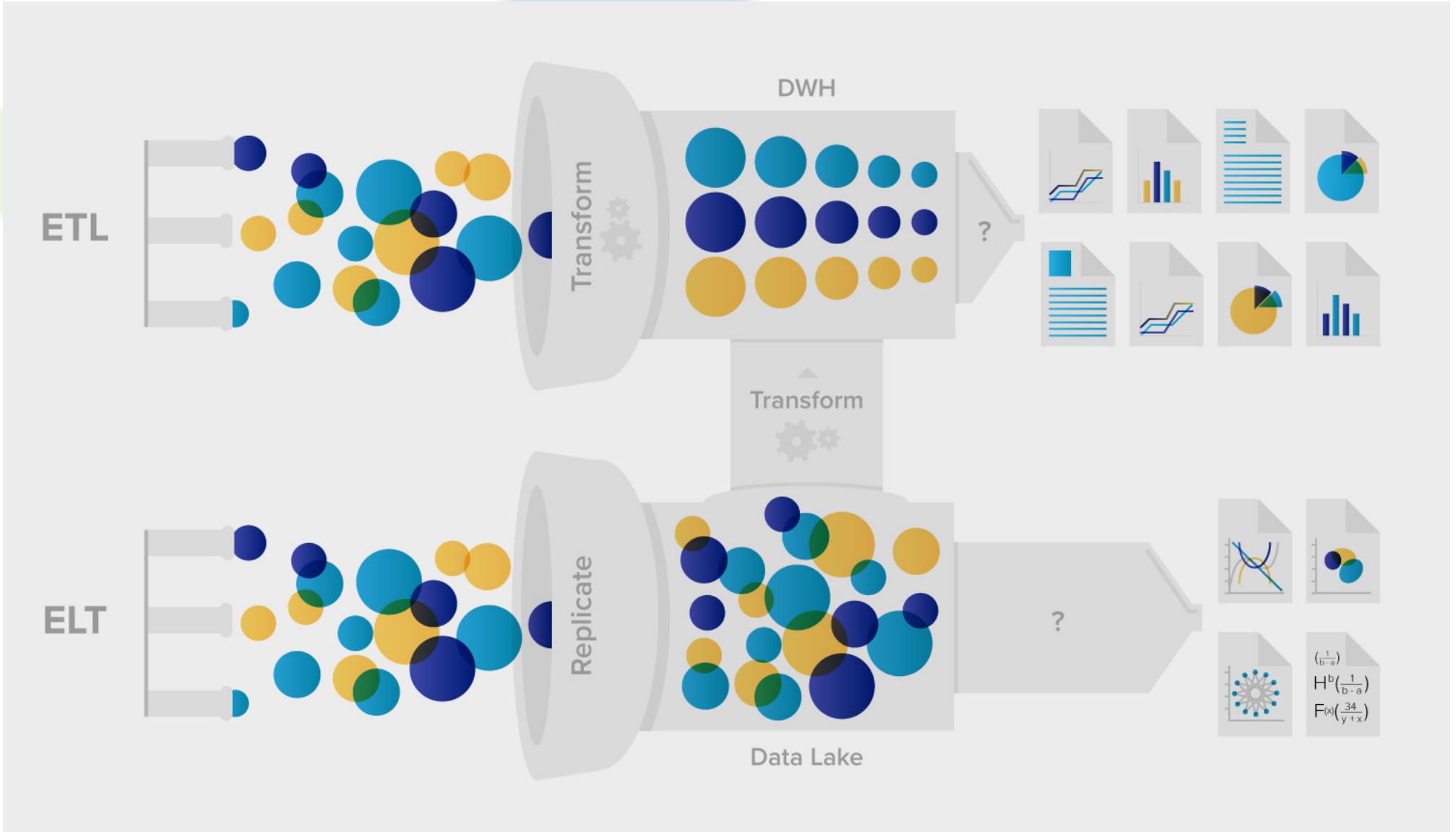
Data Science Academy



ETL x ELT

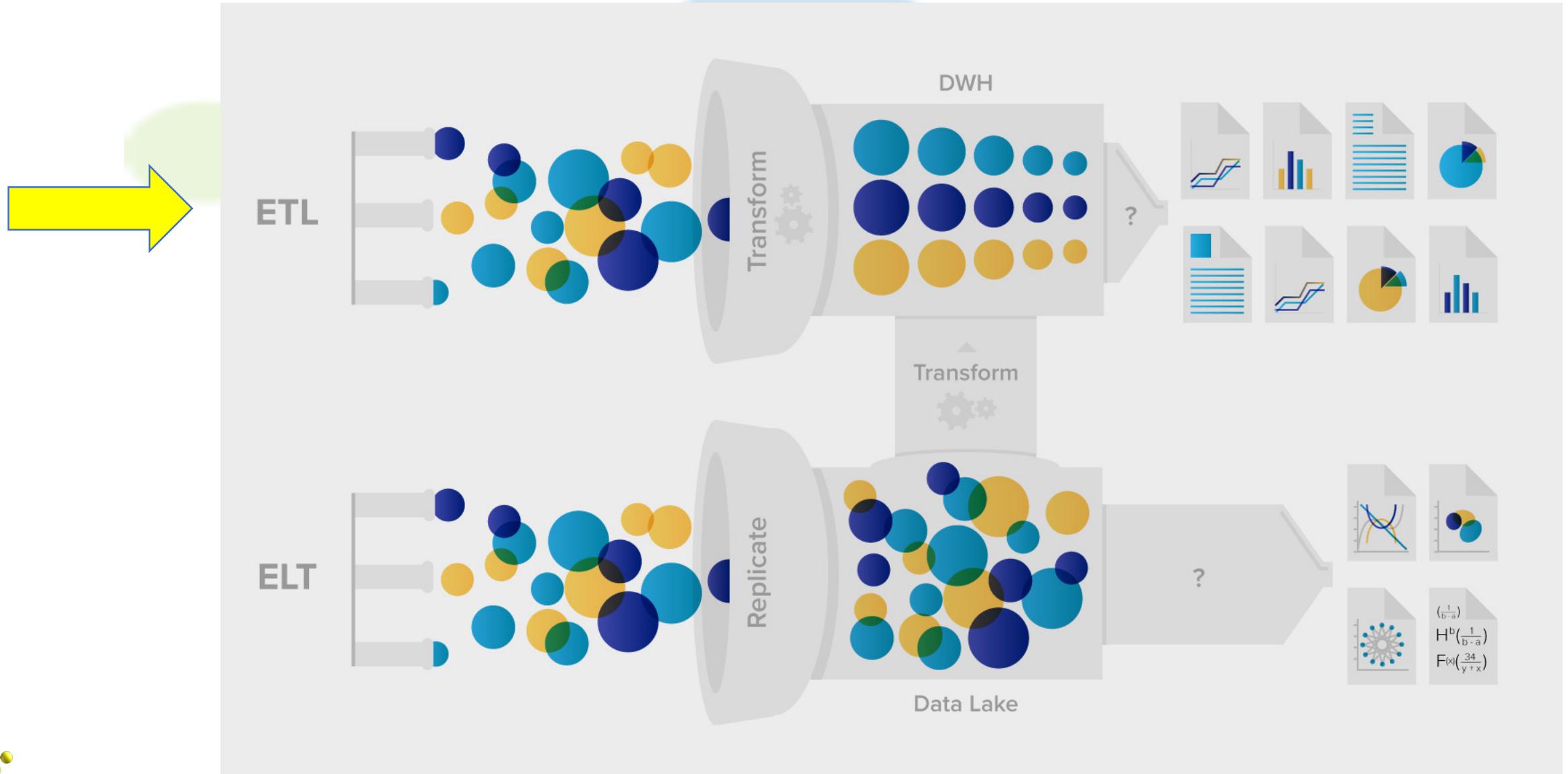
Extract,
Transform,
Load

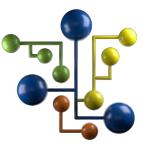
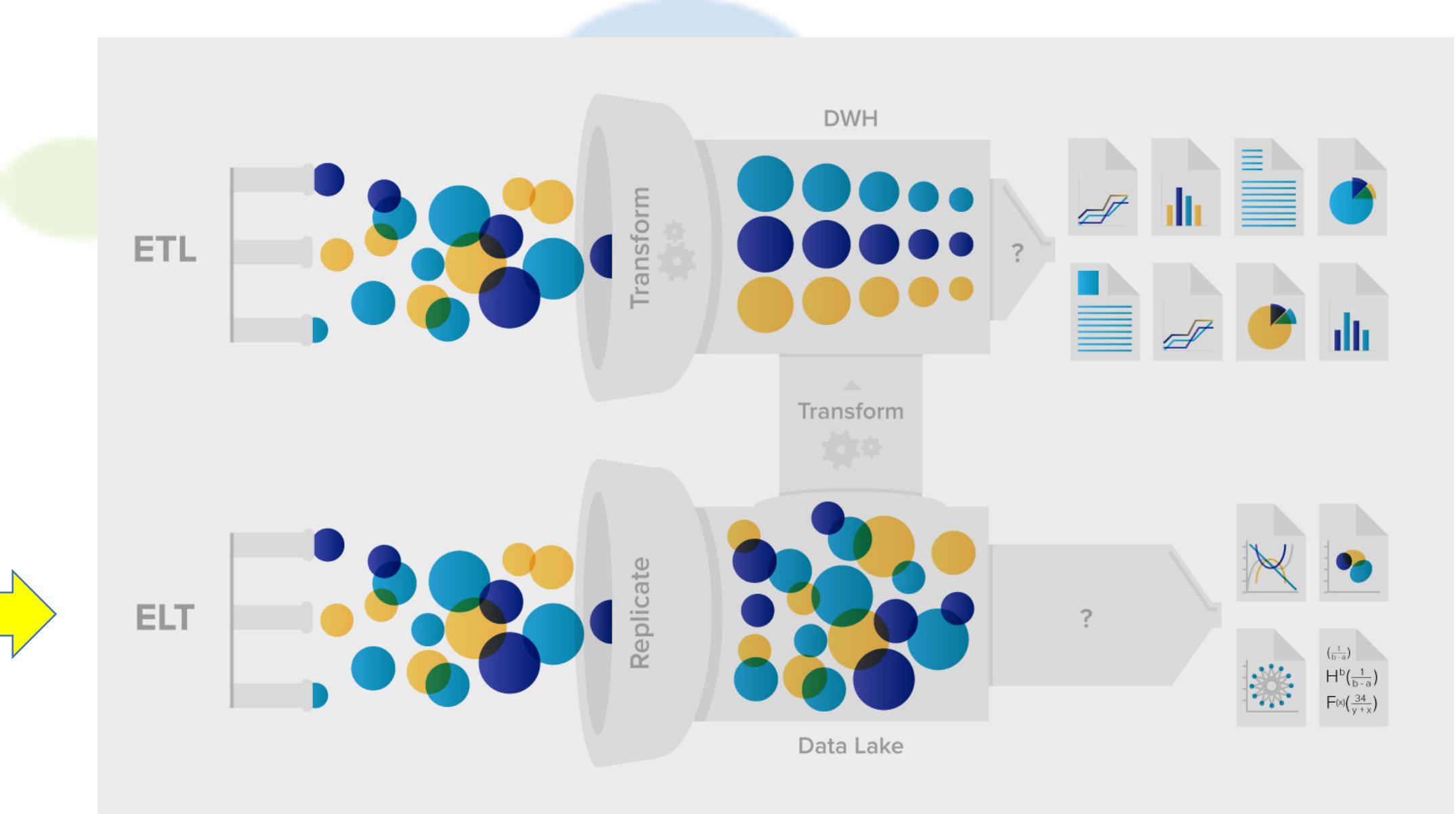
Extract,
Load,
Transform





ETL x ELT



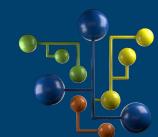
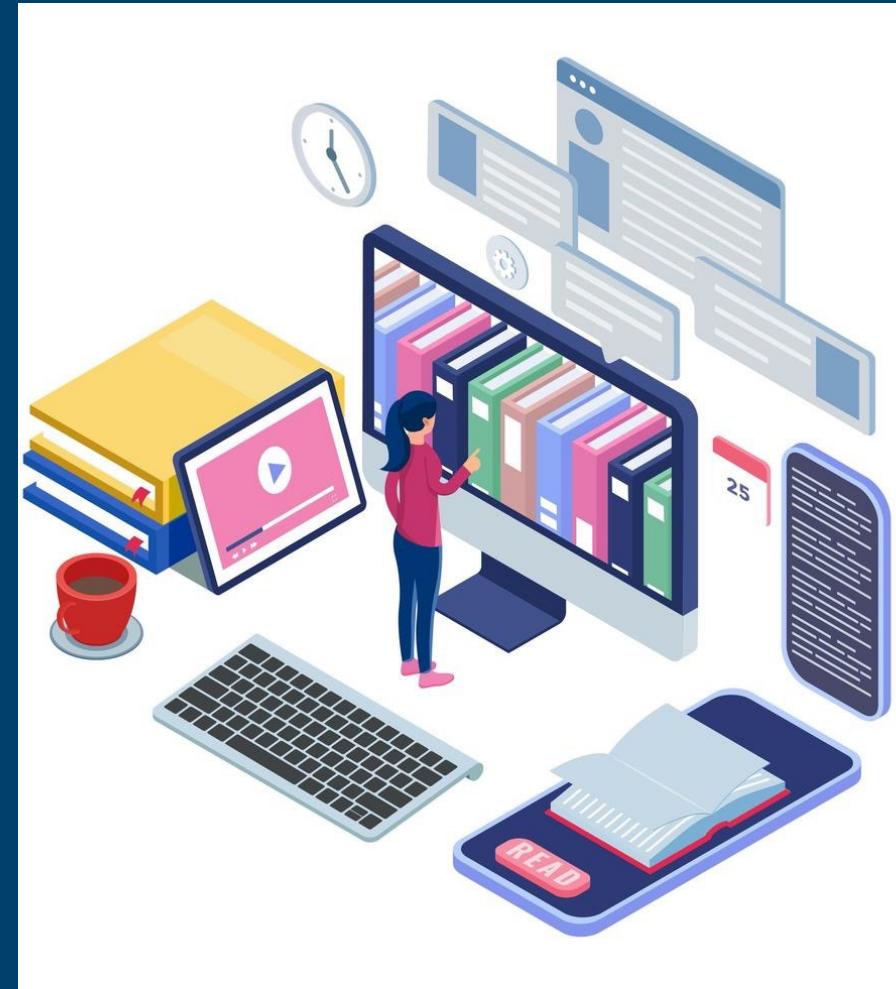
 **ETL x ELT**

Big Data Fundamentos 3.0

Como Iniciar um Projeto de Big Data?

Data Science Academy

Data Science Academy



Big Data

Fundamentos 3.0

O que é o Big Data Analytics?

Data Science Academy

Data Science Academy





O que é o Big Data Analytics?

Big Data Analytics



Data Science Academy



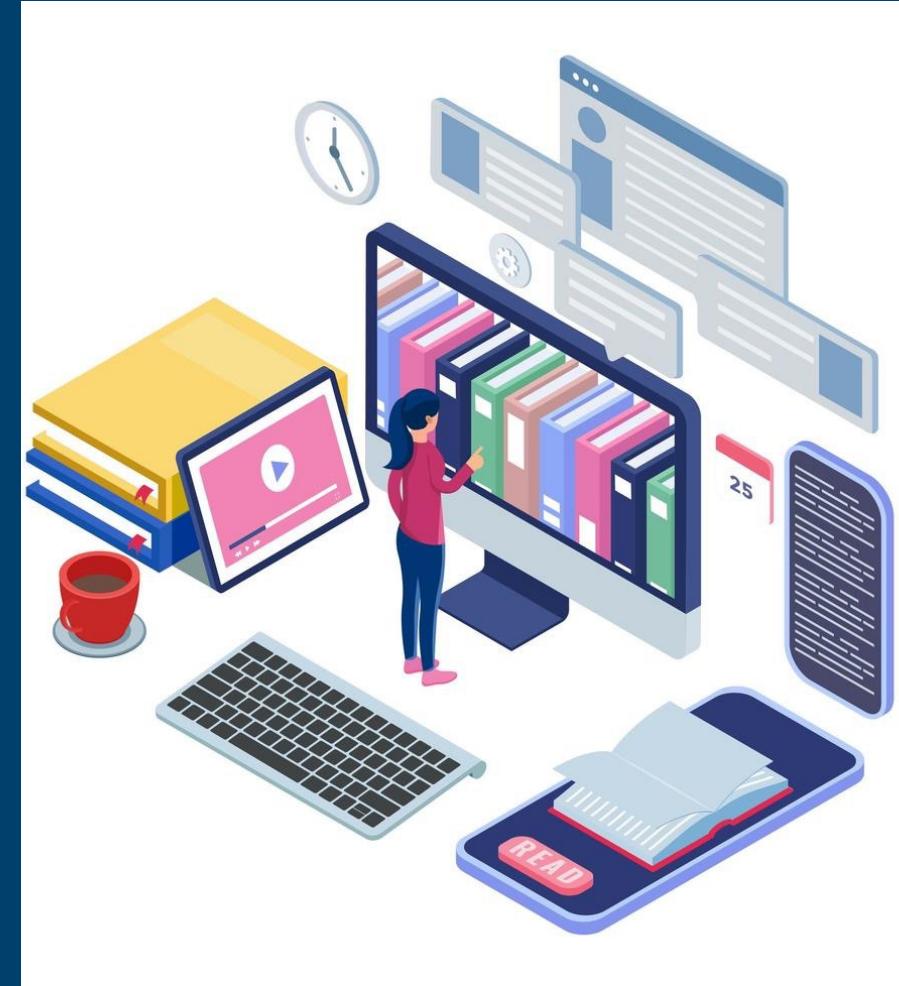


Big Data Fundamentos 3.0

Como as Empresas Estão Utilizando o Big Data?

Data Science Academy

Data Science Academy





Como as Empresas Estão Utilizando o Big Data?

Manufatura





Como as Empresas Estão Utilizando o Big Data?

Finanças





Como as Empresas Estão Utilizando o Big Data?

Saúde





Como as Empresas Estão Utilizando o Big Data?

Varejo →





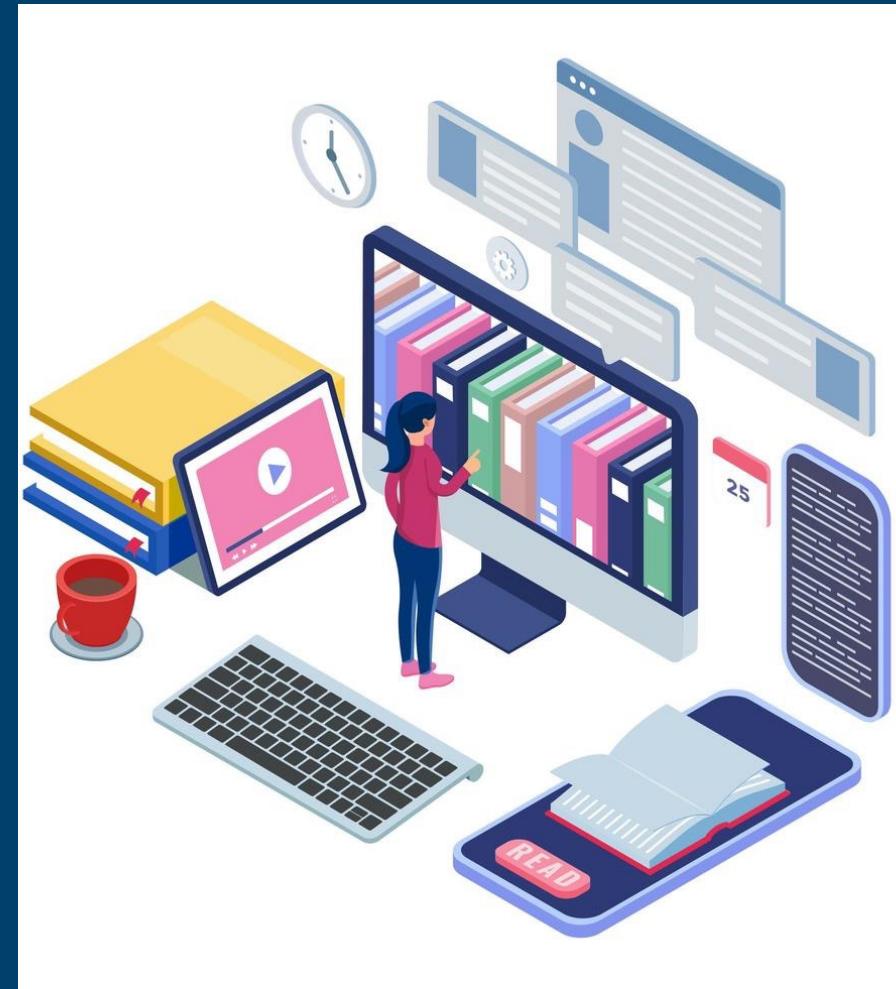
Big Data

Fundamentos 3.0

Casos de Uso de Big Data

Data Science Academy

Data Science Academy





Como as Empresas Estão Utilizando o Big Data?





Como as Empresas Estão Utilizando o Big Data?



<http://caesarscorporate.com>

A companhia de entretenimento em cassinos está usando o ambiente Hadoop para identificar diferentes segmentos de consumidor e criar campanhas de marketing específicas para cada um deles.



Como as Empresas Estão Utilizando o Big Data?



<http://caesarscorporate.com>

O novo ambiente reduziu o tempo de processamento de 6 horas para 45 minutos para posições-chave. Isso permitiu à Caesars promover uma análise de dados mais rápida e exata, aprimorando a experiência de consumidor e fazendo com que a segurança atendesse os requisitos do setor de pagamentos com cartões.



Como as Empresas Estão Utilizando o Big Data?



<http://caesarscorporate.com>

A empresa agora processa mais de 3 milhões de registros por hora.



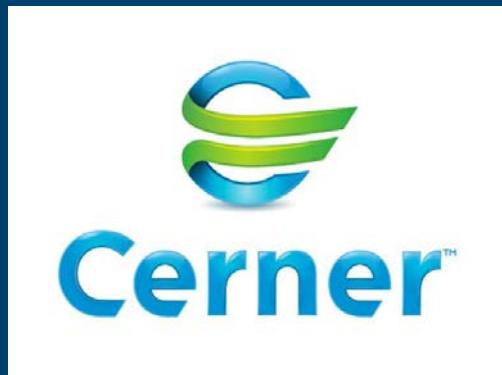
Como as Empresas Estão Utilizando o Big Data?

<http://www.cerner.com>





Como as Empresas Estão Utilizando o Big Data?

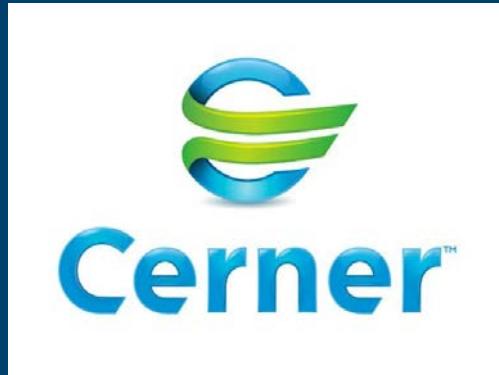


<http://www.cerner.com>

A empresa de tecnologia para o setor de saúde construiu um hub de dados corporativos no CDH (Cloudera Distribution), para criar uma visão mais compreensível de qualquer paciente, condição ou tendência.



Como as Empresas Estão Utilizando o Big Data?

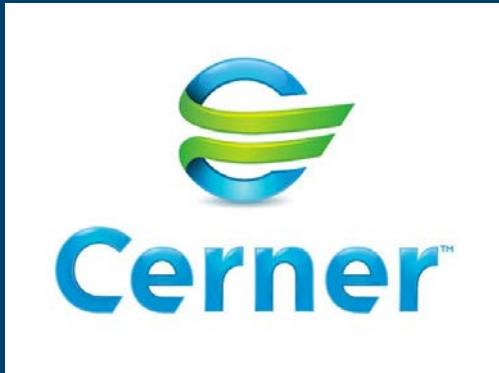


<http://www.cerner.com>

A tecnologia ajuda a Cerner e seus clientes a monitorarem mais de 1 milhão de pacientes diariamente.



Como as Empresas Estão Utilizando o Big Data?



<http://www.cerner.com>

Entre outras coisas, ela colabora na determinação mais exata da probabilidade de um paciente estar com infecção em sua corrente sanguínea.



Como as Empresas Estão Utilizando o Big Data?

eHarmony®

<http://www.eharmony.com.br>



Como as Empresas Estão Utilizando o Big Data?

eHarmony®

<http://www.eharmony.com.br>

O site de namoro online recentemente atualizou seu ambiente na nuvem, usando o CDH para analisar um volume massivo e variado de dados.



Como as Empresas Estão Utilizando o Big Data?

eHarmony®

<http://www.eharmony.com.br>

A tecnologia ajuda a eHarmony a disponibilizar novas combinações a milhões de pessoas diariamente.



Como as Empresas Estão Utilizando o Big Data?

eHarmony®

<http://www.eharmony.com.br>

O novo ambiente cloud acomoda análises mais complexas, criando resultados mais personalizados e aumentando a chance de sucesso nos relacionamentos.



Como as Empresas Estão Utilizando o Big Data?



<http://www.mastercard.com.br>



Como as Empresas Estão Utilizando o Big Data?



<http://www.mastercard.com.br>

A empresa foi a primeira a implementar a distribuição CDH do Hadoop após receber certificação PCI completa.



Como as Empresas Estão Utilizando o Big Data?



<http://www.mastercard.com.br>

A companhia usou os servidores Intel para integrar conjuntos de dados a outros ambientes já certificados.



Como as Empresas Estão Utilizando o Big Data?



<http://www.mastercard.com.br>

A MasterCard incentiva seus clientes a adotarem o sistema através do seu braço de serviços profissionais, o MasterCard Advisors.



Como as Empresas Estão Utilizando o Big Data?

FarmLogs

<https://farmlogs.com>



Como as Empresas Estão Utilizando o Big Data?

FarmLogs

<https://farmlogs.com>

A companhia de software para gerenciamento de produções agrícolas usa analytics em tempo real rodando nos processadores Intel Xeon E5 para fornecer dados sobre colheita, condições de plantio e estado da vegetação para 20% das fazendas americanas.



Como as Empresas Estão Utilizando o Big Data?

FarmLogs

<https://farmlogs.com>

A tecnologia ajuda os fazendeiros a aumentarem a produtividade de seus acres.



Como as Empresas Estão Utilizando o Big Data?



<http://www.nipponpaint.com>



Como as Empresas Estão Utilizando o Big Data?



<http://www.nipponpaint.com>

Uma das maiores fornecedoras de tinta da Ásia usa os processadores Intel Xeon E7 v2 para compreender o comportamento de clientes, otimizar sua cadeia de suprimentos e melhorar suas campanhas de marketing.



Como as Empresas Estão Utilizando o Big Data?



<http://www.nipponpaint.com>

A Nippon Paint agora testa um novo sistema baseado no Hadoop para usufruir das ferramentas de alto desempenho e processar Big Data.



Como as Empresas Estão Utilizando o Big Data?

Outras empresas usando Hadoop:

Empresa	Especificações Técnicas	Utilização
Facebook	Mais de 12 TB de storage	Hadoop é utilizado em soluções de relatórios e Machine Learning
Twitter	--	Hadoop é usado desde 2010 para o processamento de logs e tweets
LinkedIn	4100 nodes Hadoop	Todos os dados do LinkedIn passam através de um cluster Hadoop
Yahoo!	4500 nodes Hadoop e mais de 1 TB de storage	Usado no portal do Yahoo
Ebay	4000 nodes Hadoop	Um dos maiores clusters Hadoop que se tem notícia, usado para processar as mais de 300 milhões de pesquisas feitas pelos usuários



Como as Empresas Estão Utilizando o Big Data?

Outras empresas usando Hadoop:

Empresa	Especificações Técnicas	Utilização
Accenture	De acordo com a demanda do cliente	Projetos de Big Data na área financeira, telecom e varejo
Ning	--	Plataforma de Rede Social, utiliza o Hadoop para relatórios e Big Data Analytics
Spotify	690 nodes em cluster Hadoop, totalizando 38 TB de memória RAM e 28 PB de storage	Usa Hadoop para geração de conteúdo e agregação de dados
Fox	70 nodes Hadoop	Usado para análise de logs e Machine Learning



Como as Empresas Estão Utilizando o Big Data?





Big Data Fundamentos 3.0

Como Iniciar um Projeto de Big Data?

Data Science Academy



Data Science Academy





Como Iniciar um Projeto de Big Data?



1. Definição do Business Case
2. Planejamento do Projeto
3. Definição dos Requisitos Técnicos
4. Criação de um “Total Business Value Assessment”

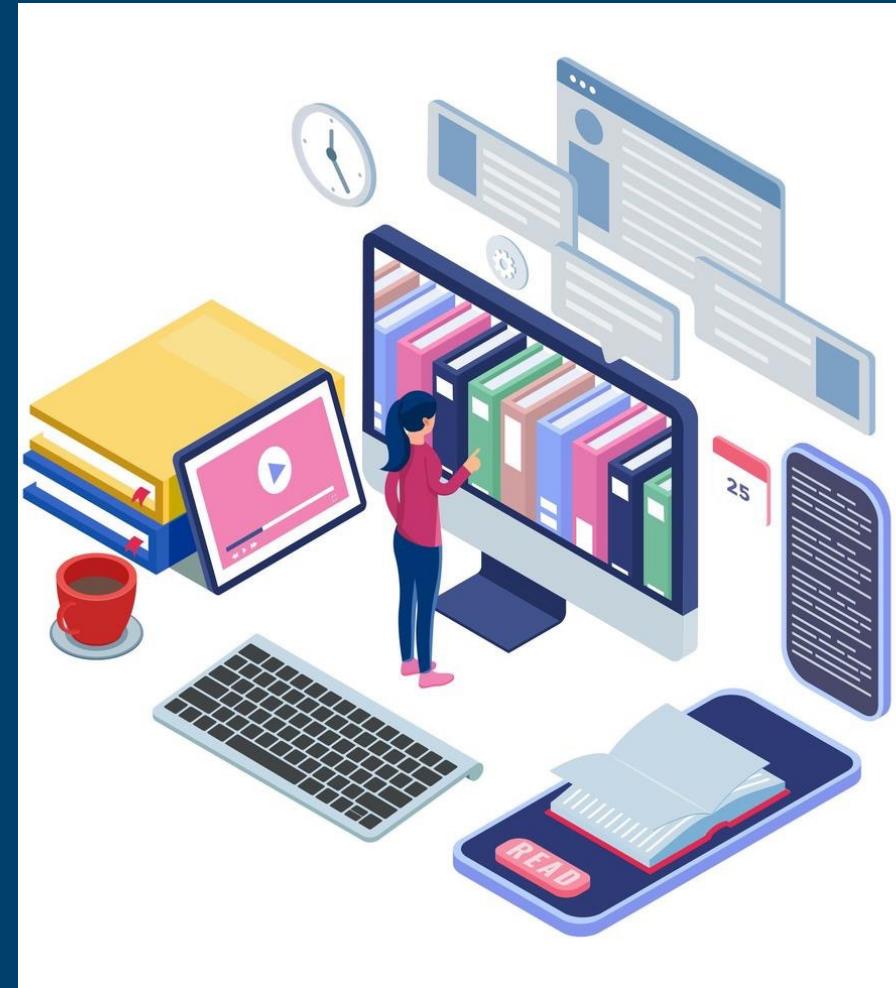
Big Data

Fundamentos 3.0

Encerramento

Data Science Academy

Data Science Academy





Obrigado Por Acompanhar Este Curso!

Palavra Final do Instrutor

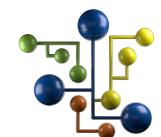
Por Onde Eu Começo?

Avaliação Final

3 tentativas, 50 questões, 120 minutos

Certificado de Conclusão

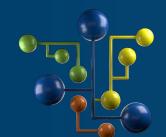
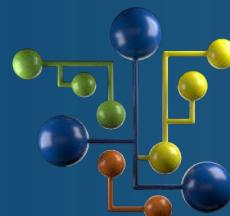
E-Book do Curso



Muito Obrigado!

Tenha Uma Excelente
Jornada de Aprendizado.

Data Science Academy



Data Science Academy