

BIG **Data**

Fundamentos 2.0



Equipe Data Science Academy

Esse e-book não pode ser usado para fins comerciais, mas pode ser distribuído livremente sob a licença [Creative Commons](#).

Pedimos apenas, a gentileza de citar a fonte, pois todo este material é resultado de trabalho árduo de nossa equipe.

Esta foi a forma que encontramos de contribuir com a sociedade que deve ter a educação como prioridade.





Data Science
Academy

Big Data Fundamentos 2.0

O futuro é aqui.



Data Science Academy

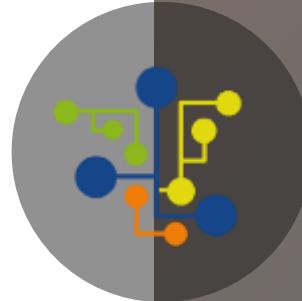


Data Science Academy

Prezado(a) Aluno(a),

Seja muito bem-vindo(a) ao curso:

Big Data Fundamentos 2.0



- › Apresentação do Instrutor
- › Apresentação da DSA
- › Conteúdo do Curso
- › Por que Realizar Este Curso?
- › Benefícios Deste Curso

Treinamentos Gratuitos DSA

Acreditamos que aprender não para nunca.

Introdução à Ciência de Dados

Big Data Fundamentos 2.0

Python Fundamentos Para Análise de Dados

Microsoft Power BI Para Data Science

Nossa Equipe

preparou esses treinamentos
especialmente
para você!

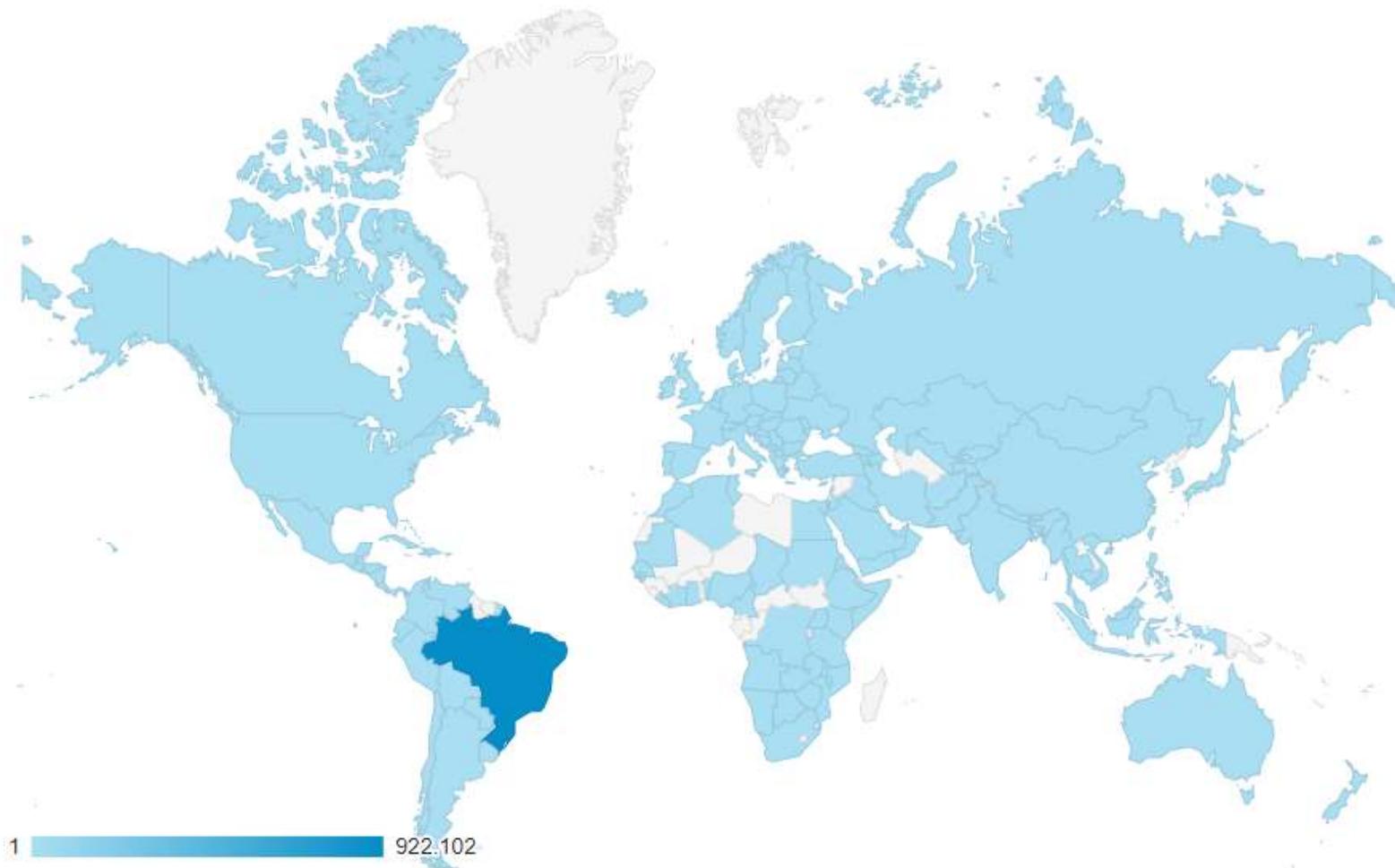
Data Science Academy

A Data Science Academy é um portal de ensino online especializado em Big Data, Machine Learning, Inteligência Artificial, Desenvolvimento de Chatbots 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

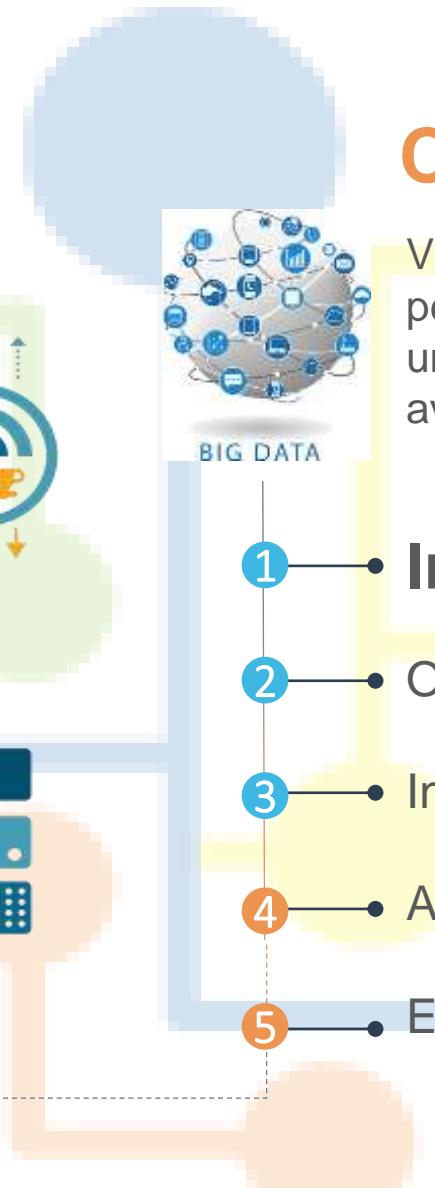
Data Science Academy – Localização

No Brasil e no Mundo.



Conteúdo Programático

Seja Muito Bem-Vindo(a)!



Conteúdo do Curso

Visão geral de conceitos e definições que permitam uma compreensão clara do que é o universo do Big Data para que você possa avançar sua carreira nesta vibrante área.

Introdução

- 1 • O que é Big Data?
- 2 • Introdução ao Hadoop
- 3 • Arquitetura Hadoop
- 4 • Ecossistema Hadoop

Conteúdo Programático

Seja Muito Bem-Vindo(a)!



Conteúdo do Curso

Visão geral de conceitos e definições que permitam uma compreensão clara do que é o universo do Big Data para que você possa avançar sua carreira nesta vibrante área.

- 6 • Soluções Comerciais com Hadoop
- 7 • Introdução ao Apache Spark
- 8 • Banco de Dados NoSQL
- 9 • Como Iniciar Projetos de Big Data
- 10 • **Avaliação e Certificado de Conclusão**

Curso Big Data Fundamentos 2.0

Objetivo

Este curso oferece uma introdução detalhada dos principais conceitos envolvendo Big Data, permitindo uma compreensão clara do que há de mais avançado em tecnologia de Engenharia de Dados.

Carreira

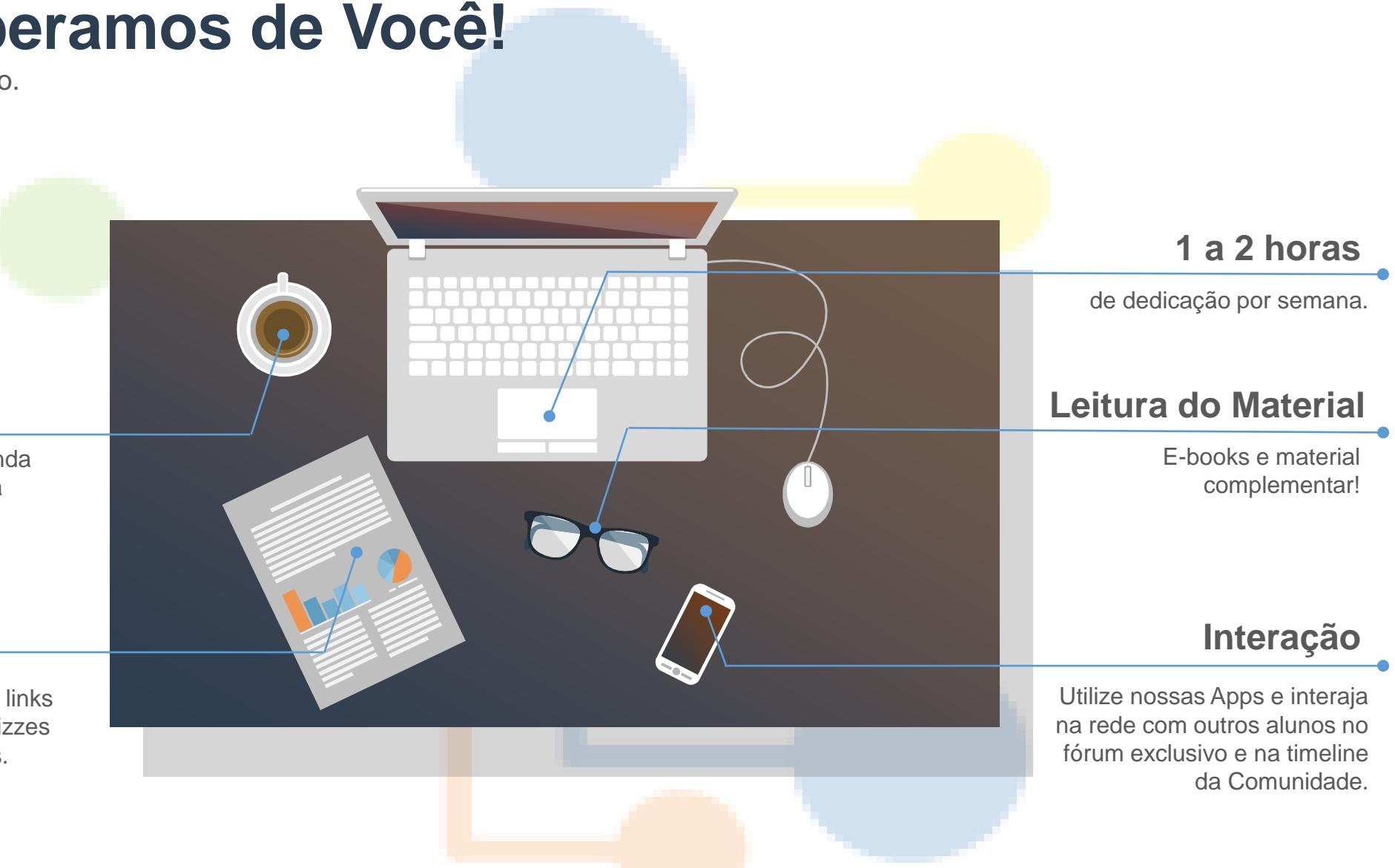
Compreensão clara do que é o universo do Big Data para que você possa avançar sua carreira nesta vibrante área.

Pré-Requisitos



O Que Esperamos de Você!

Sua abordagem no curso.



Objetivos ao concluir o curso Big Data Fundamentos 2.0

Crescimento do Big Data

Apache Hadoop

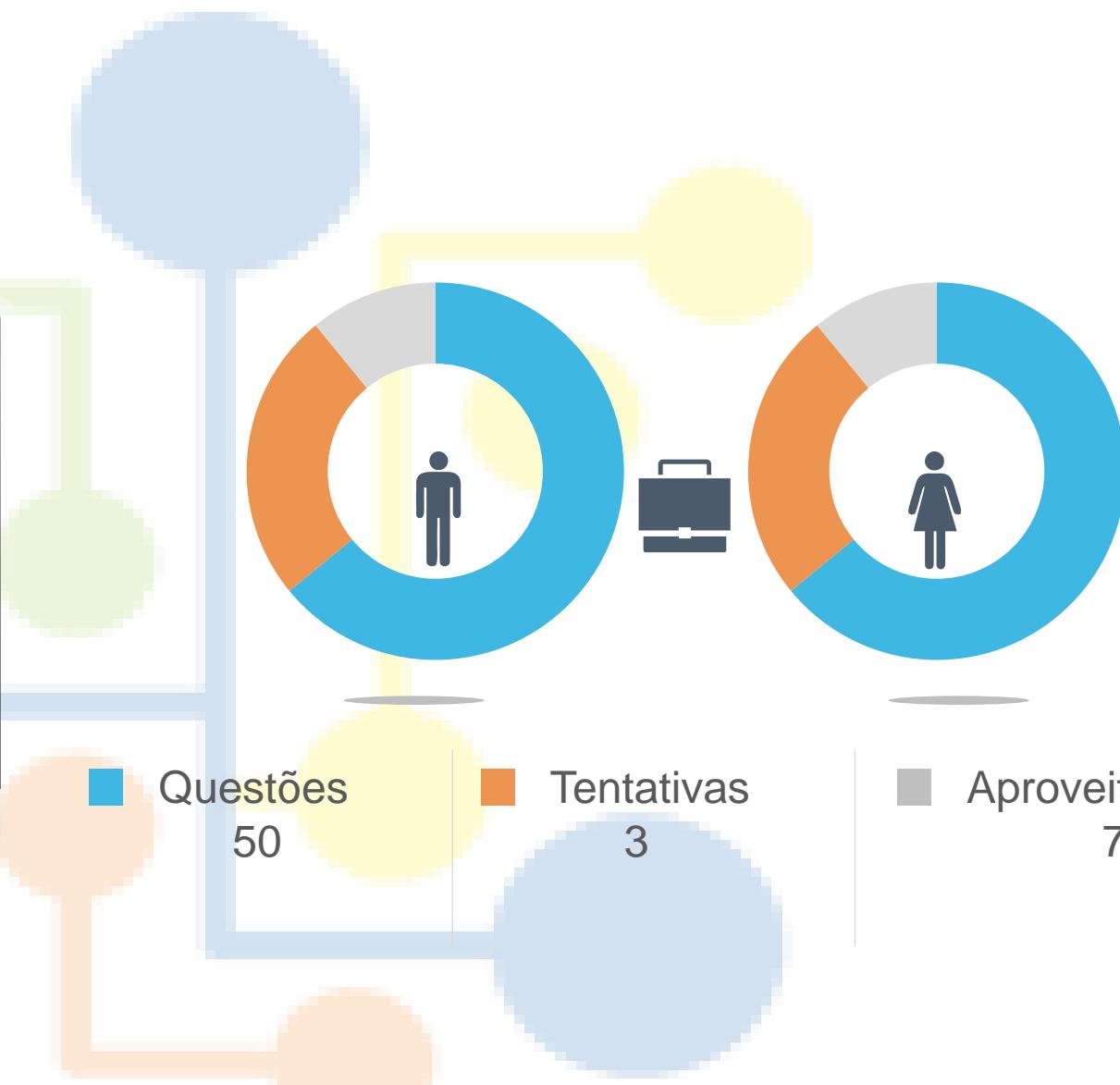
Déficit de Profissionais

Compreensão do Big Data

Apache Spark

Alavancar a Sua Carreira

Avaliação Final



➤ Suporte



Em todos os nossos cursos, gratuitos e pagos, o aluno recebe suporte em até 24 horas, incluindo finais de semana e feriados.

Utilize um dos nossos canais de comunicação e obtenha suporte sempre que precisar!

Nossa equipe é obcecada pelo sucesso dos nossos alunos!

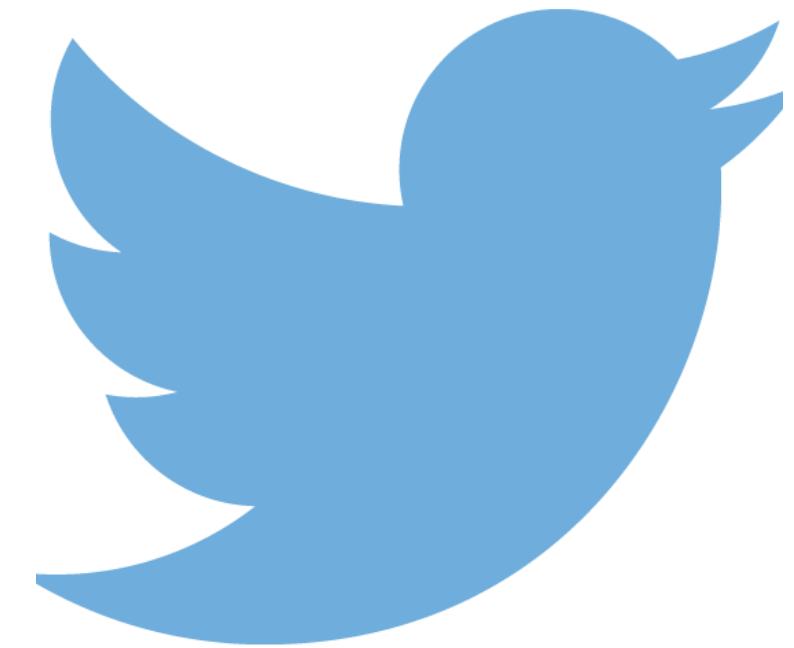
Oferecemos mais do que cursos online.
Oferecemos uma experiência de aprendizagem!



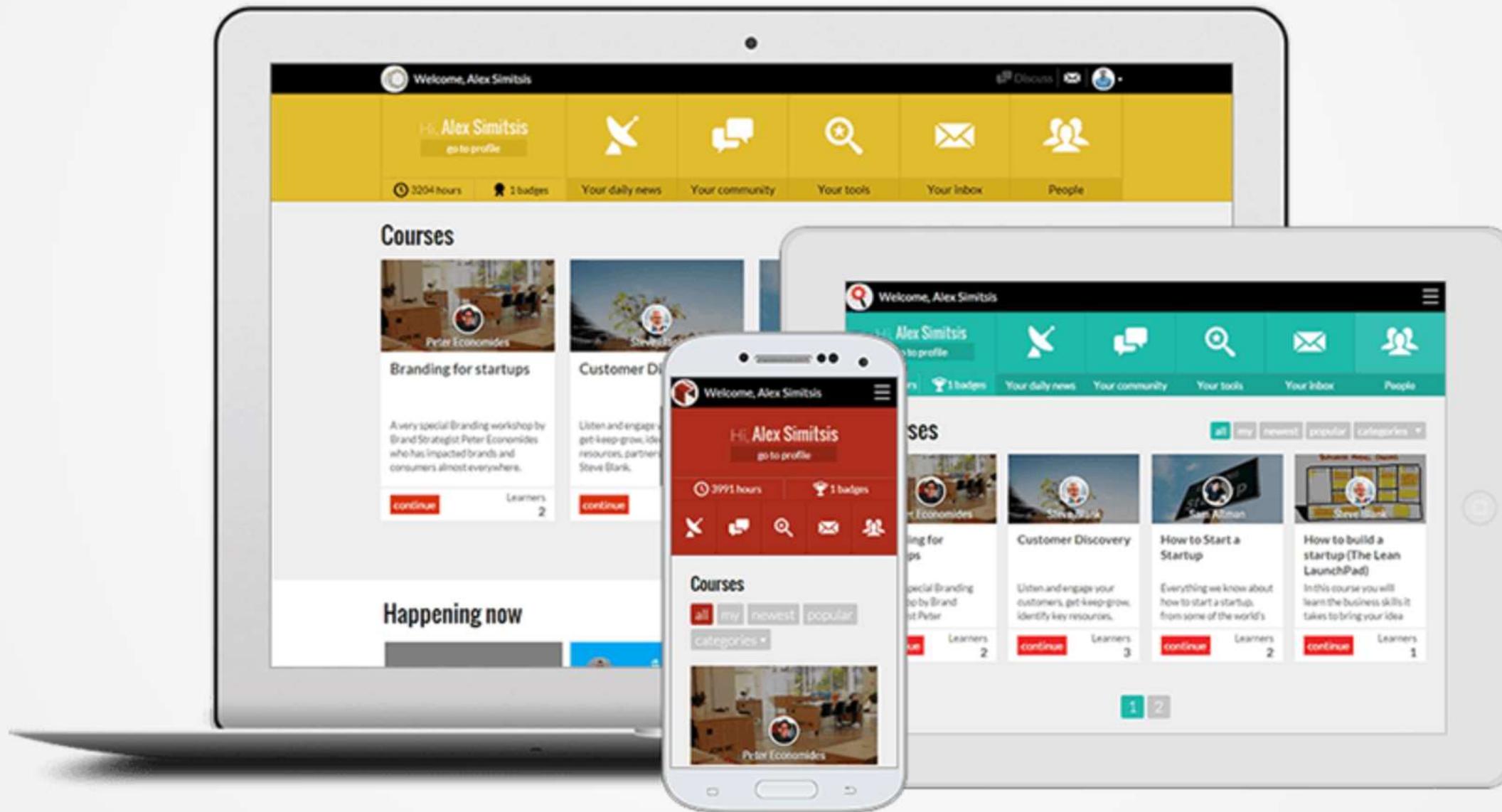
Seja aprovado,

E receba o e-book com todo o
conteúdo do curso
de forma gratuita.

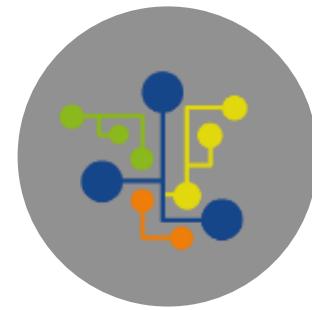
Compartilhe seu Certificado de Conclusão



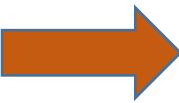
Apps Gratuitas para iOS e Android



Trilhas de Aprendizagem



Formações DSA



Formação Cientista de Dados
Formação IA
Formação Java



Formação Engenheiro de Dados



Formação

Cientista de Dados



Formação

Engenheiro de Dados



Formação

Inteligência Artificial



Formação

Java para Data Science



Formação Cientista de Dados

Transforme Dados em Resultados



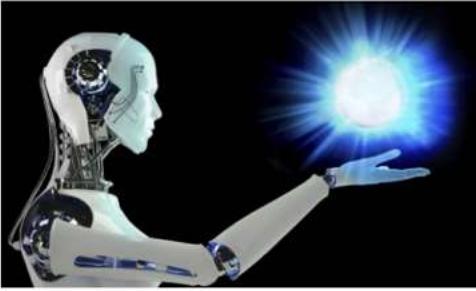
Formação Cientista de Dados

1. Big Data Analytics com R e Microsoft Azure Machine Learning
2. Big Data Real-Time Analytics com Python e Apache Spark
3. Engenharia de Dados com Hadoop e Spark
4. Machine Learning com R e Python
5. Business Analytics
6. Visualização de Dados e Design de Dashboards
7. Preparação para Carreira de Cientista de Dados





Formação Inteligência Artificial



Formação Inteligência Artificial

1. Introdução à Inteligência Artificial
2. Deep Learning Frameworks
3. Programação Paralela em GPU
4. Deep Learning I
5. Deep Learning II
6. Visão Computacional e Reconhecimento de Imagens
7. Processamento de Linguagem Natural e Reconhecimento de Voz
8. Análise em Grafos Para Big Data
9. Sistemas Cognitivos
10. Projeto – Assistente Virtual Inteligente





Formação

Java para Data Science



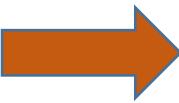
1. Java Fundamentos
2. Análise Preditiva com Machine Learning em Java
3. Aplicações de Inteligência Artificial com Deep Learning em Java
4. Projeto – Aplicação Analítica Mobile com Inteligência Artificial

Formação

Java para Data Science



Formações DSA



Formação Cientista de Dados
Formação IA
Formação Java



Formação Engenheiro de Dados



Formação Engenheiro de Dados



1. Design e Implementação de Data Warehouses
2. Data Lake - Design, Projeto e Integração
3. Segurança e Alta Disponibilidade de Dados
4. Machine Learning e IA em Ambientes Distribuídos
5. Analytics - Visualização, Relatórios e Tomada de Decisões com Big Data

Formação

Engenheiro de Dados



SQL Para Big Data
Gerenciamento de Dados com MongoDB
Arquiteto de Soluções AWS – Preparação Para Certificação

Data Mining e Modelagem Preditiva
Machine Learning com Linguagem Scala e Apache Spark
Desenvolvimento de Chatbots
R Fundamentos Para Análise de Dados



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.

Big Data

Modelos de análise de dados estruturados, possuem limitações quando precisam tratar grandes volumes de dados.



Muitas empresas não sabem que dados precisam ser analisados

Muitas empresas nem mesmo sabem que os dados estão disponíveis

Dados preciosos são descartados por falta de conhecimento ou ferramentas de tratamento





**É caro manter e
organizar
grandes volumes
de dados não-
estruturados**





Big Data

Estamos em um período de transformação no modo em que dirigimos nossos negócios e, principalmente, as nossas vidas.



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.



Big Data

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



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 roadmaps cada vez mais elaborados, apontando e, até prevendo, o comportamento de empresas e clientes.





Big Data

Big Data

Entre 2005 e 2020, o universo digital irá crescer de 130 exabytes para 40.000 exabytes ou 40 trilhões de gigabytes

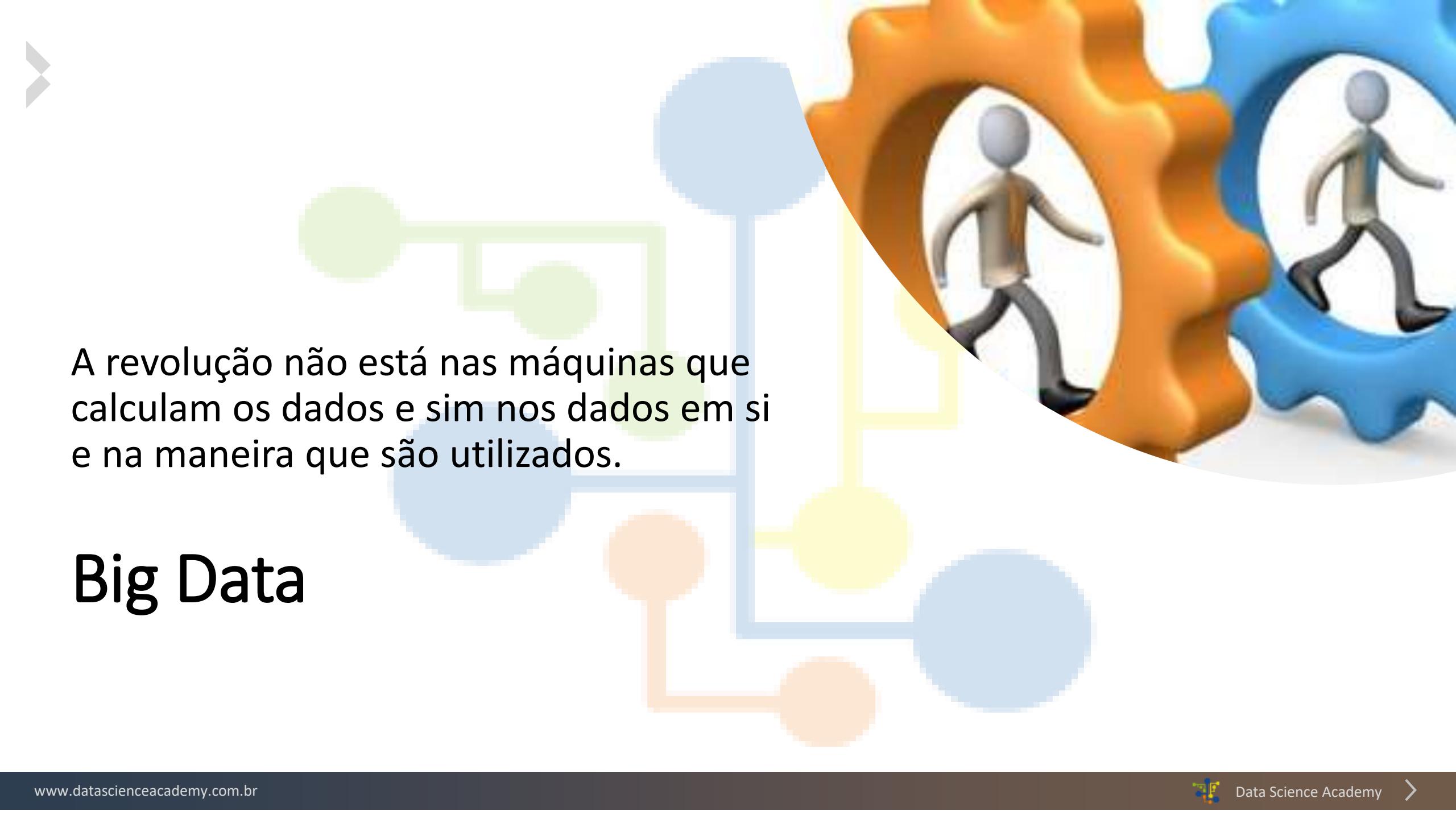
Em 2020, haverá 5.200 gigabytes para cada homem, mulher e criança no planeta

Até 2020, o universo digital irá dobrar de tamanho a cada 2 anos





Dados – Matéria-prima dos negócios



A revolução não está nas máquinas que calculam os dados e sim nos dados em si e na maneira que são utilizados.

Big Data

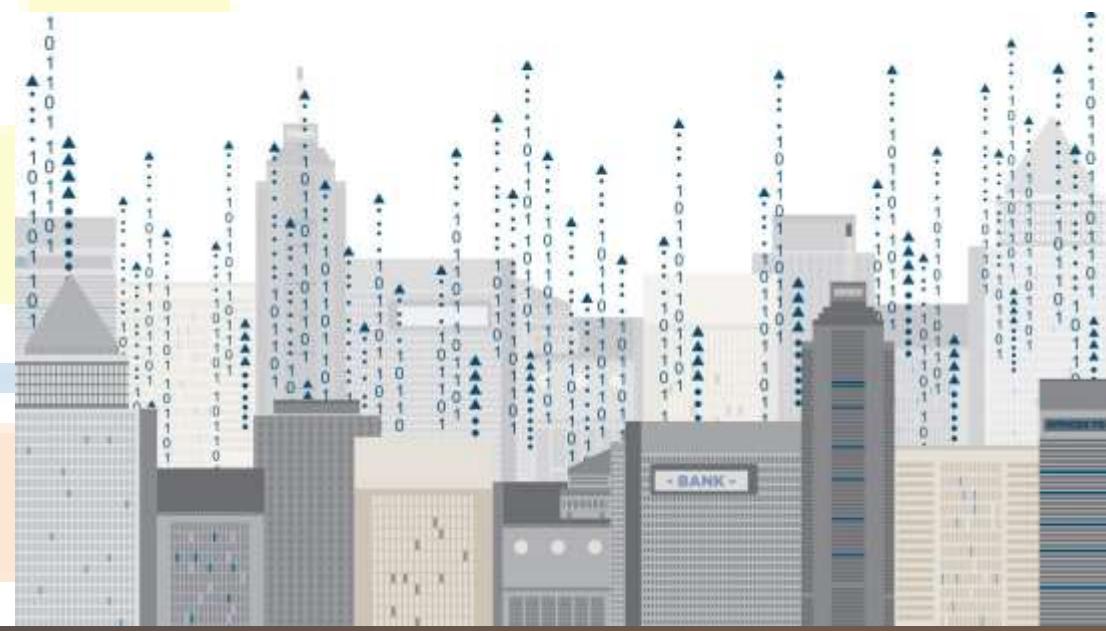




Mas afinal, o que é Big Data?

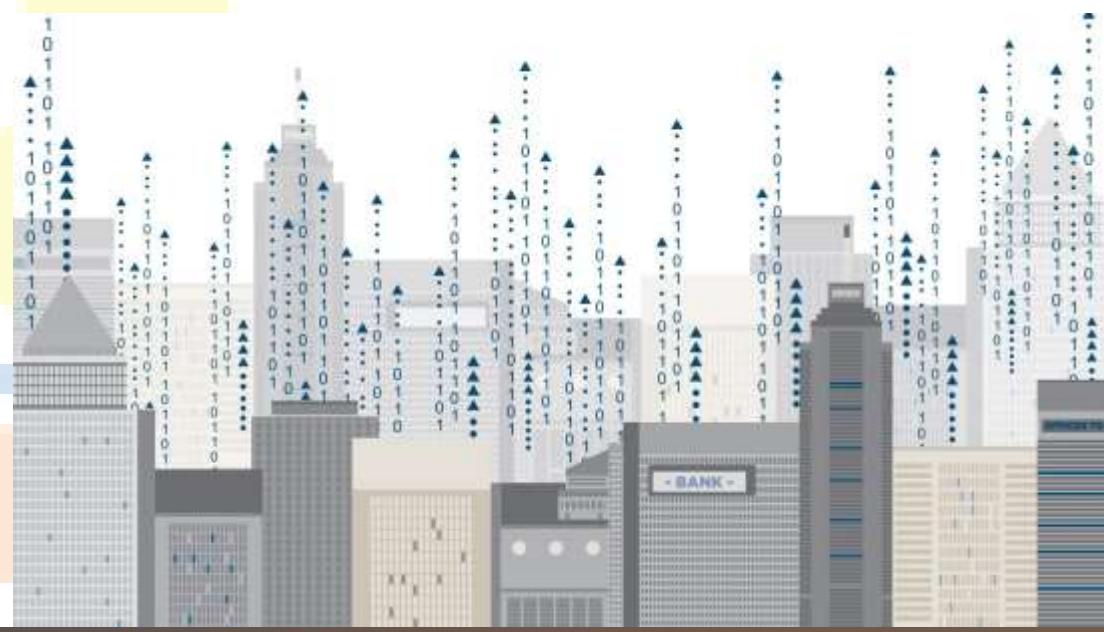
O que é Big Data?

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?

Capacidade de uma sociedade de obter informações de maneiras novas a fim de gerar ideias úteis e bens e serviços de valor significativo.

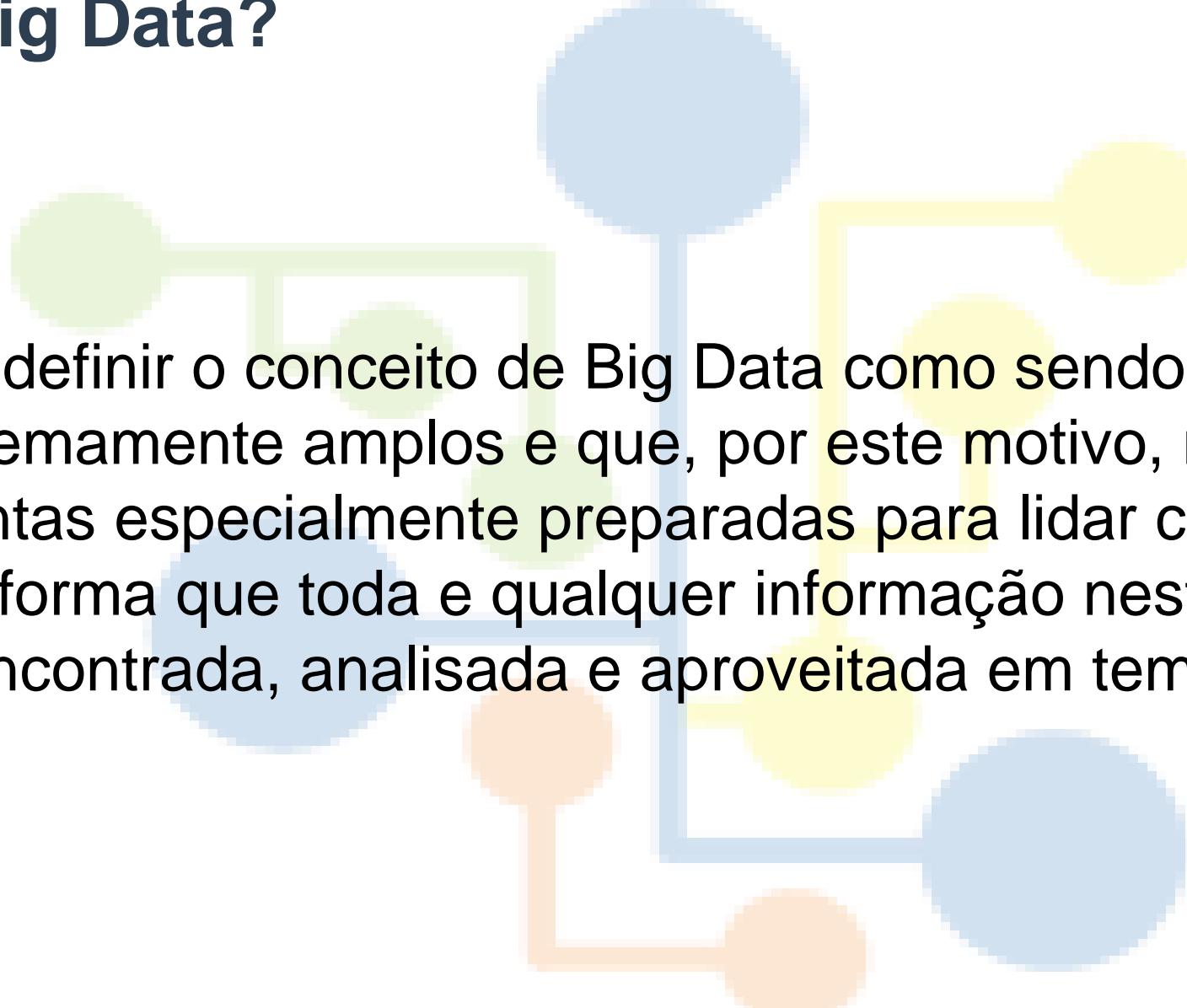


O que é Big Data?

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.

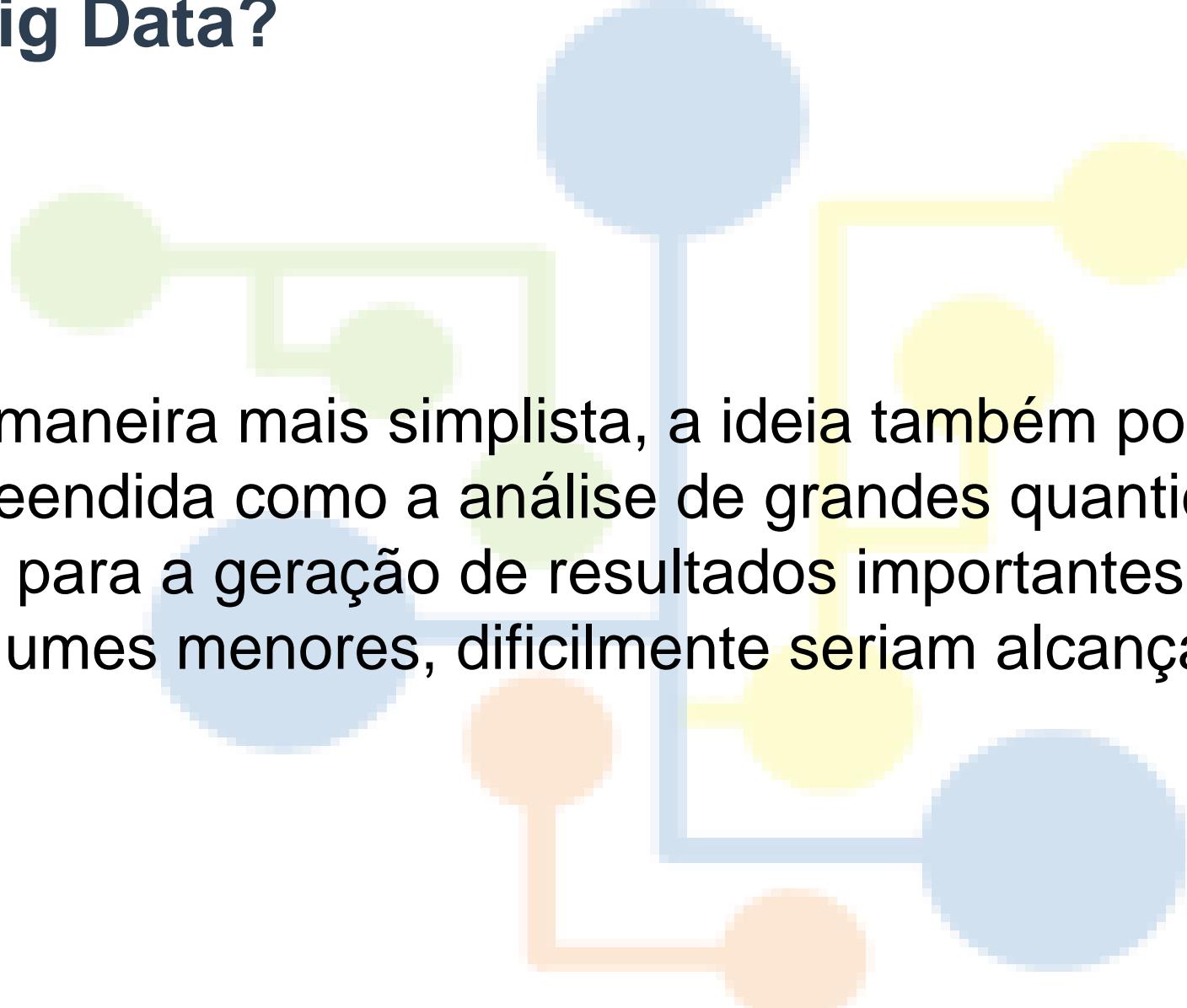


O que é Big Data?



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, de forma que toda e qualquer informação nestes meios possa ser encontrada, analisada e aproveitada em tempo hábil.

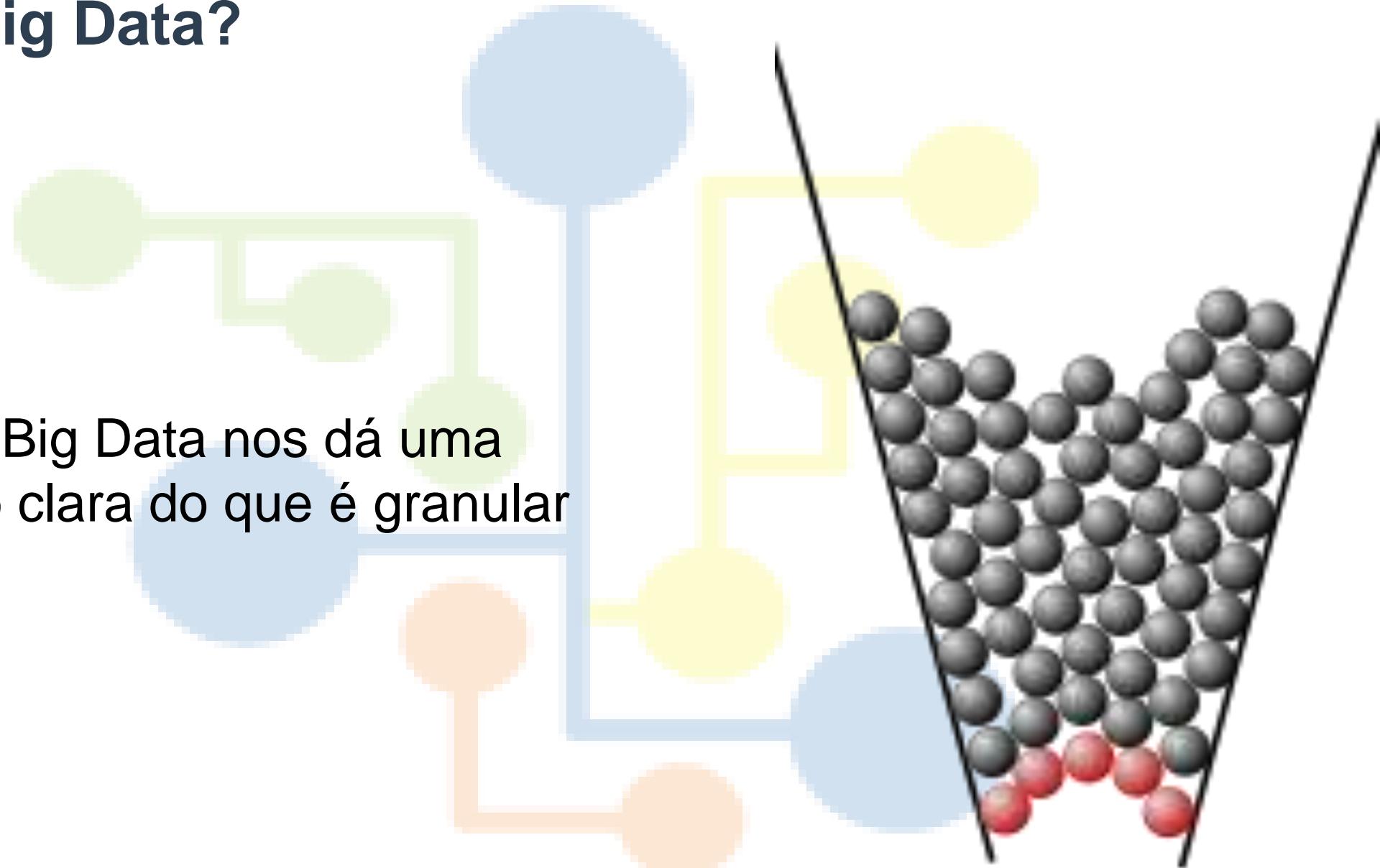
O que é Big Data?



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.

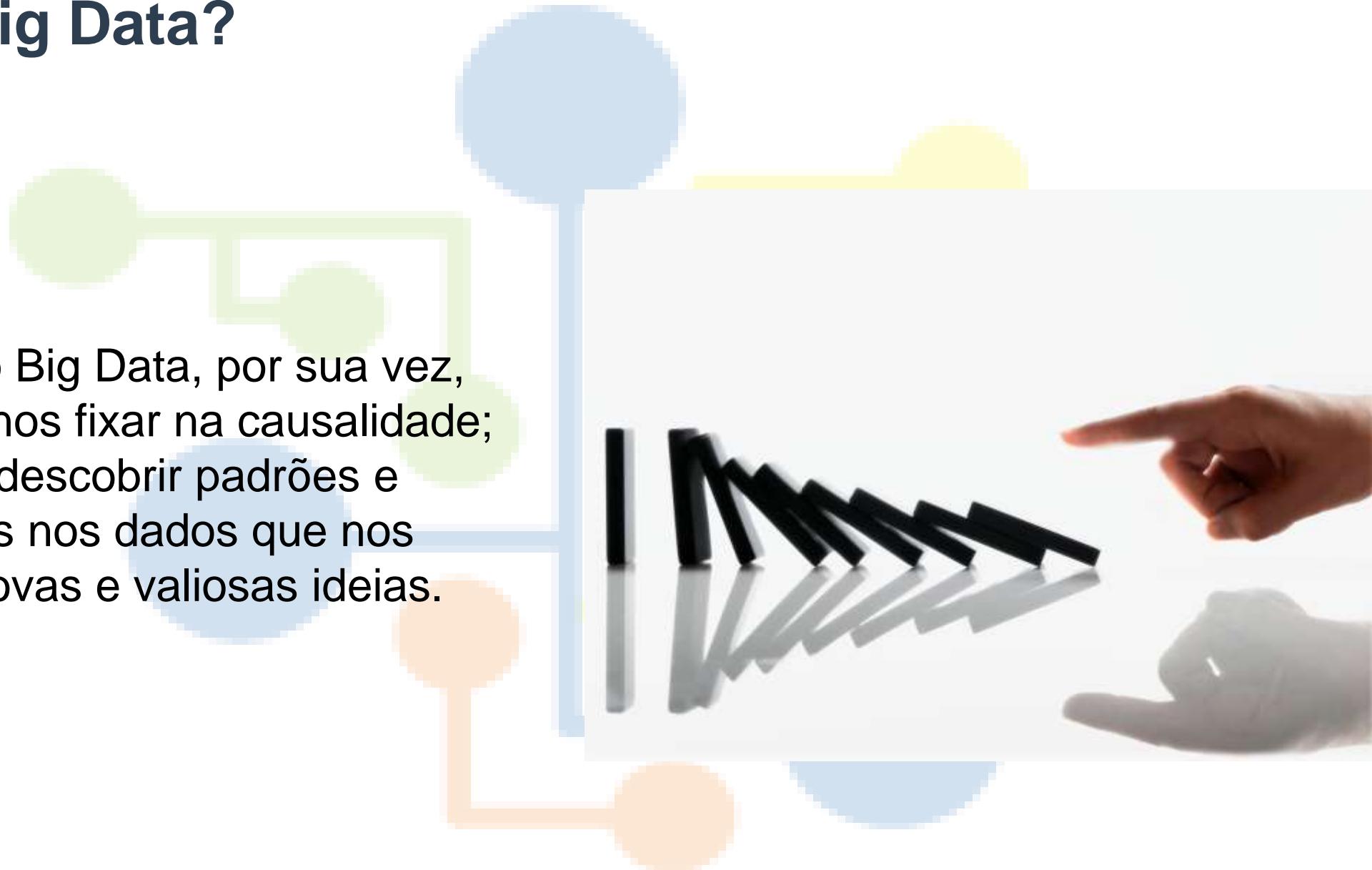
O que é Big Data?

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



O que é Big Data?

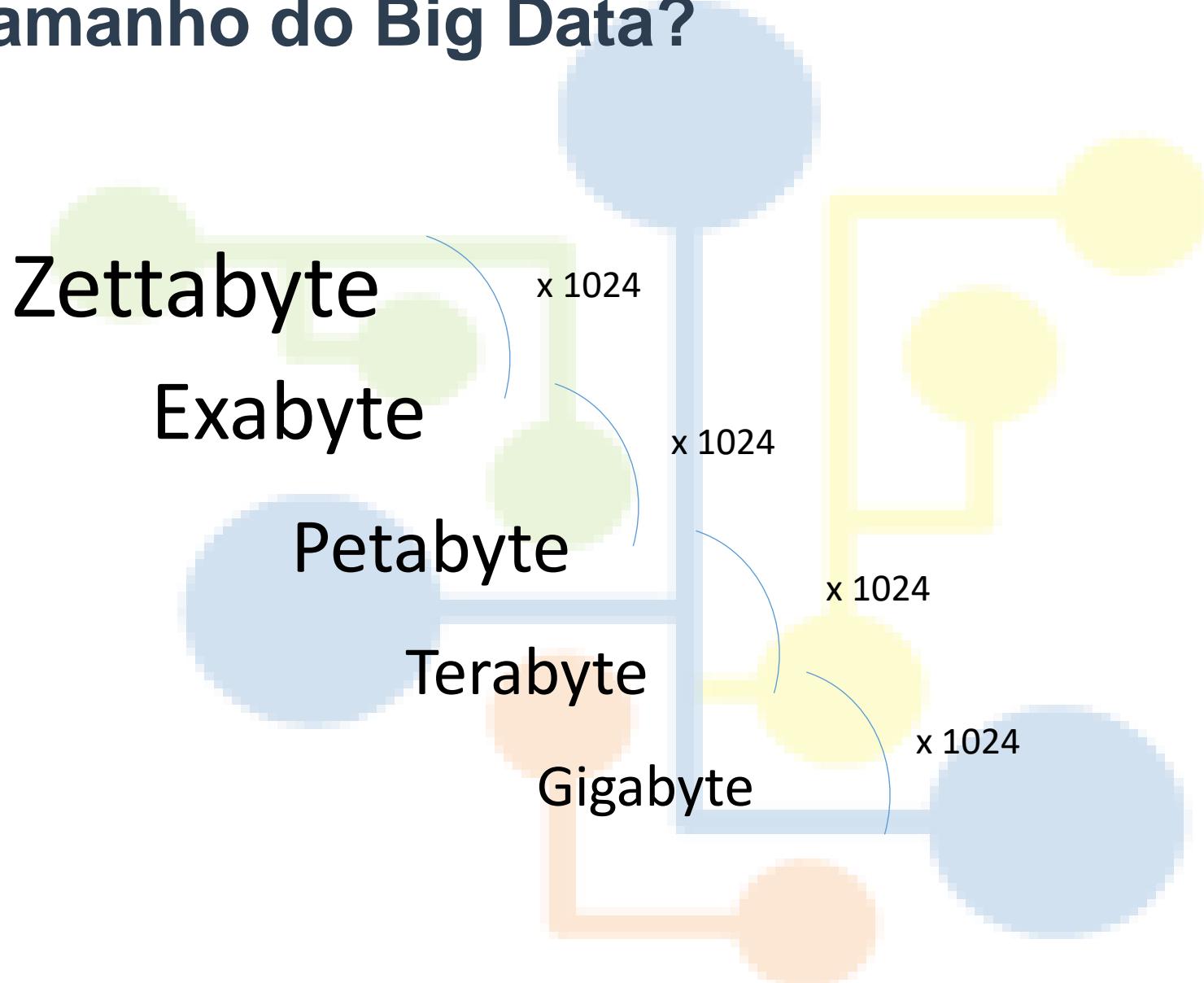
No mundo do Big Data, por sua vez,
não temos de nos fixar na causalidade;
podemos descobrir padrões e
correlações nos dados que nos
propiciem novas e valiosas ideias.



Qual o Tamanho do Big Data?



Qual o Tamanho do Big Data?



Qual o Tamanho do Big Data?

2017 *This Is What Happens In An Internet Minute*



O que é Big Data?

Muitos dos dados gerados, possuem um tempo de vida curto e se não analisados, perdem a utilidade.

Dados são transformados em informação, que precisa ser colocada em contexto para que possa fazer sentido.

É caro integrar grandes volumes de dados não-estruturados.



O que é Big Data?

Dados potencialmente valiosos em sistemas ERP, CRM ou SCM são descartados ou perdidos apenas porque ninguém presta atenção neles.



A hand is shown from the side, palm facing up, holding a glowing blue and white globe. The globe is surrounded by a complex network of glowing white lines and dots, representing a global communication or data network. The background is a dark blue gradient.

A Importância do Big Data

Qual a Importância do Big Data?



Porque surgiram tecnologias que permitem processar esta grande quantidade de dados de forma eficiente e com baixo custo

E por que Big Data tem se tornado tão importante?

Qual a Importância do Big Data?



Os dados podem ser analisados em seu formato nativo, seja ele estruturado, não estruturado ou streaming (fluxo constante de dados)

E por que Big Data tem se tornado tão importante?

Qual a Importância do Big Data?



Dados podem ser
capturados em tempo real

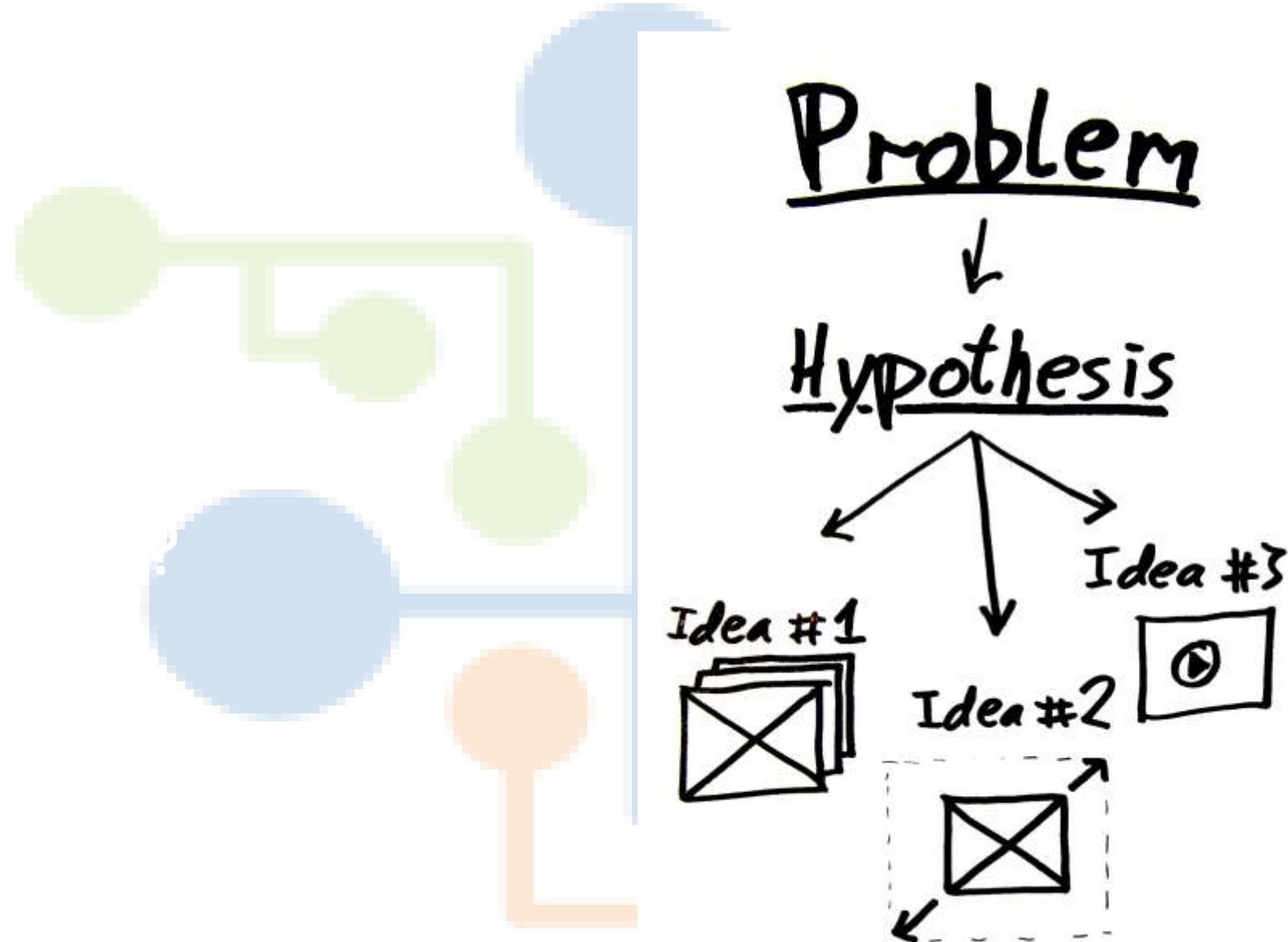
E por que Big Data tem se
tornado tão importante?

Qual a Importância do Big Data?



Dados podem ser transformados em insights de negócios

E por que Big Data tem se tornado tão importante?



Qual a Importância do Big Data?

Desafios

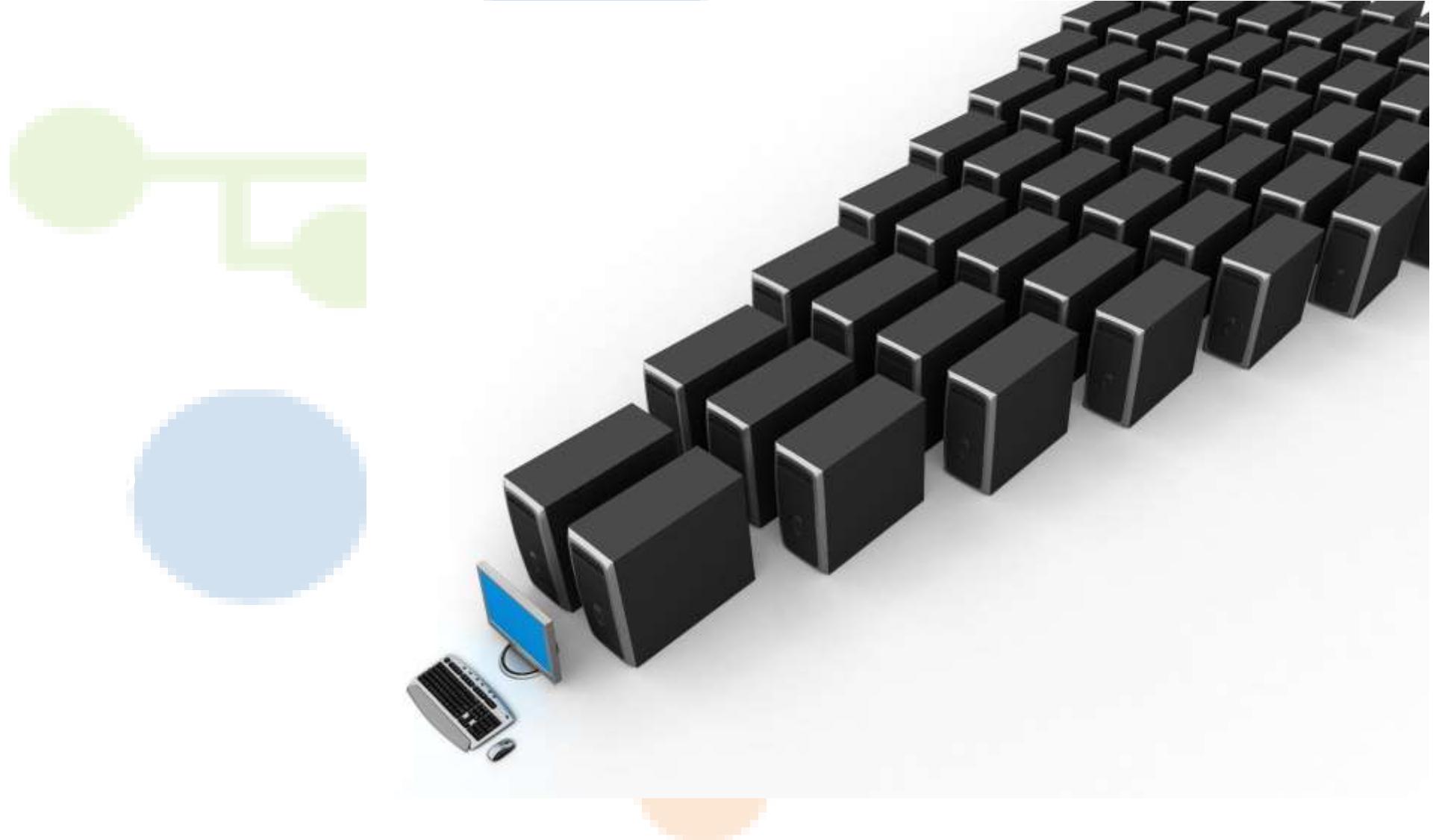
Encontrar profissionais habilitados em Big Data.

Compreender as plataformas e ferramentas para Big Data.

Coletar, armazenar e analisar dados de diferentes fontes, em diferentes formatos e gerados em diferentes velocidades.

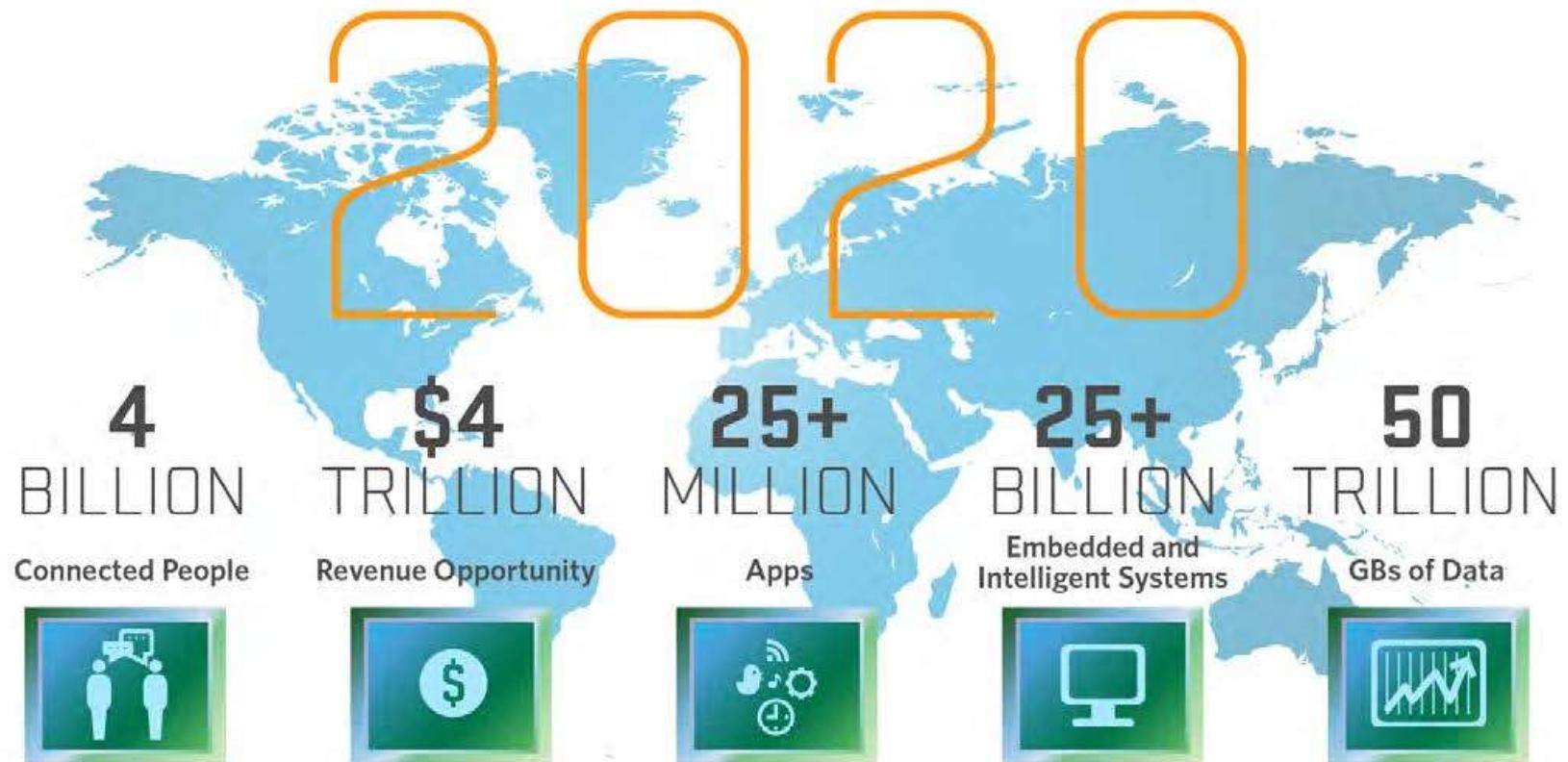
Migrar do sistema tradicional de coleta e armazenamento de dados, para uma estrutura de Big Data.

Qual a Importância do Big Data?





Portância de



Source: Mario Morales, IDC

Qual a Importância do Big Data?

BIG DATA LANDSCAPE 2017



Last updated 4/5/2013

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark (@firstmarkcap)

mattturk.com/bigdata2017

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

A Importância do Big Data

Até 2018, haverá um deficit de 140 a 190 mil profissionais com habilidades em análise de dados e mais de 1,5 milhão de gerentes e analistas que saibam usar Big Data de forma efetiva para tomada de decisões.

McKinsey Global Institute "Big Data Report 2015"

Os 4 V's do Big Data

Volume

Tamanho dos Dados.

Variedade

Formato dos Dados

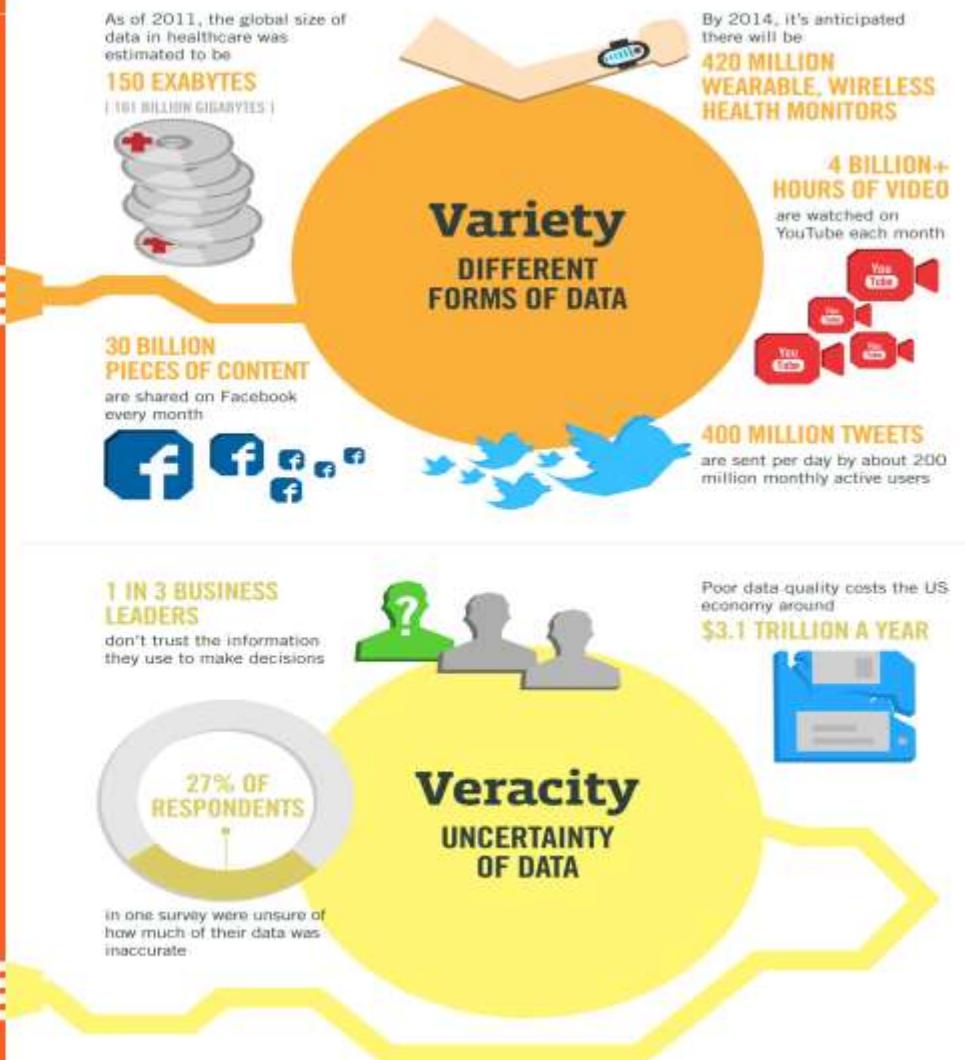
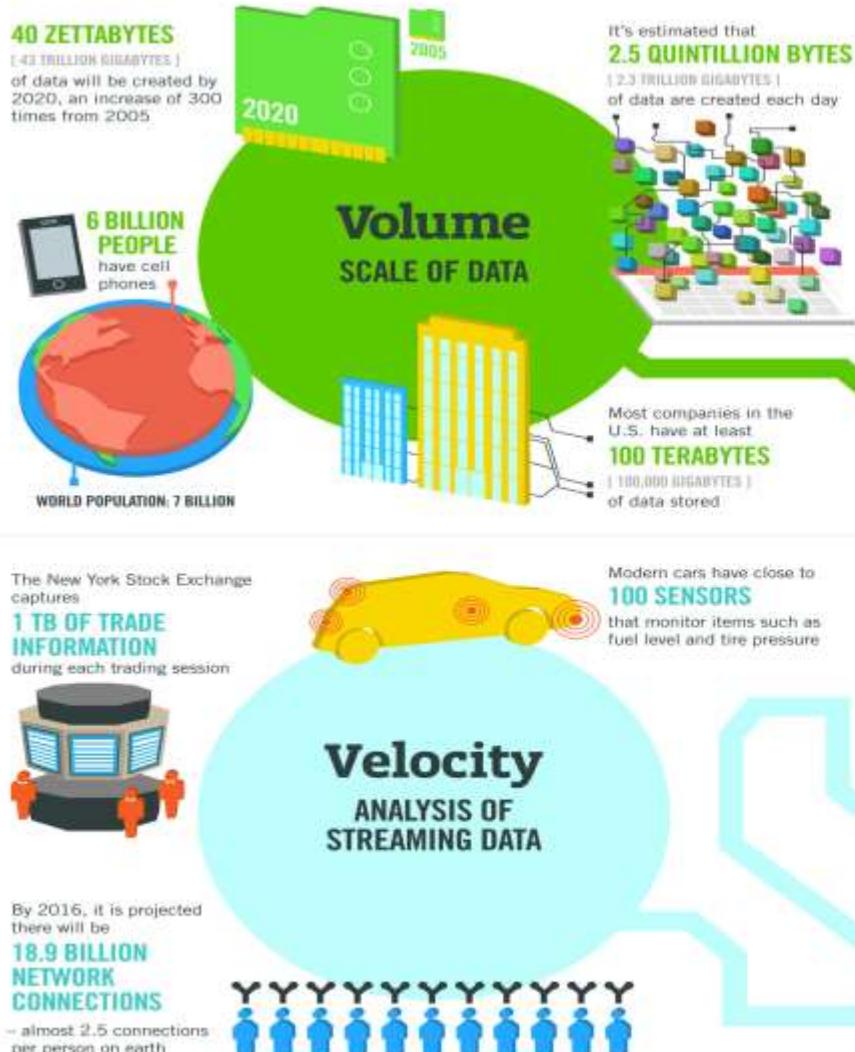
Velocidade

Geração dos Dados.

Veracidade

Confiabilidade dos Dados

Os 4Vs do Big Data



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



Os 4Vs do Big Data

Volume

Tamanho dos Dados.

- Espera-se que 40 zettabytes de dados sejam criados até 2020 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 4Vs 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 no ano de 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 4Vs 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 existirão até 2016.

Os 4Vs do Big Data

Veracidade

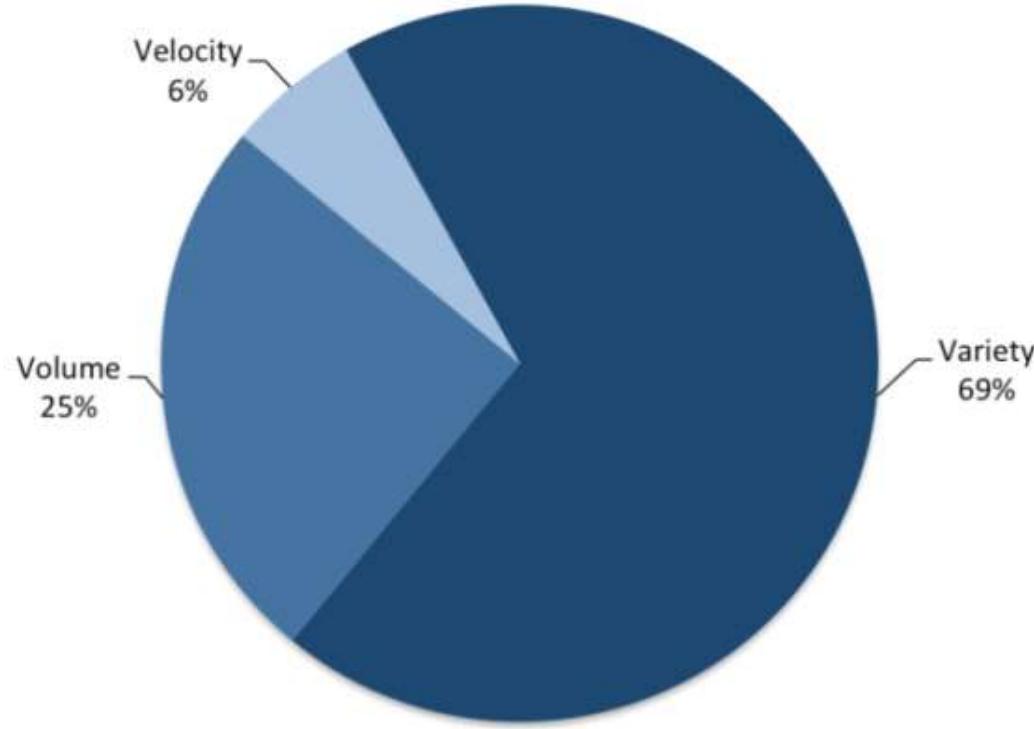
confiabilidade dos Dados.

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.

Os 4Vs do Big Data

Importância: Volume, Velocidade, Variedade



Os 4Vs do Big Data



O Big Data traz um oceano de oportunidades!

Os 4Vs do Big Data

Processar de forma eficiente e com baixo custo grandes volumes de dados

Transformar 12 TB de tweets gerados cada dia em produtos de análise de sentimento



Responder ao aumento da velocidade de geração dos dados

Investigar 5 milhões de eventos de trade nas bolsas de valores a fim de identificar fraudes



Coletar e analisar dados de diferentes formatos e fontes

Monitorar milhares de vídeos de segurança a fim de identificar pontos perigosos em uma cidade

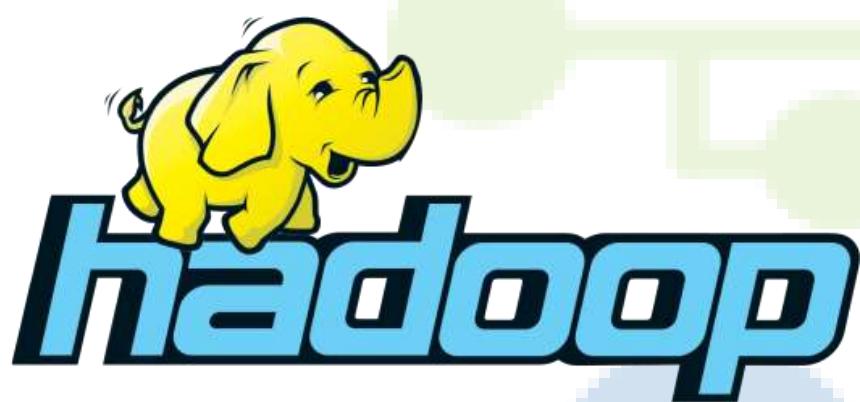


Garantir que os dados sejam confiáveis



Introdução ao Hadoop

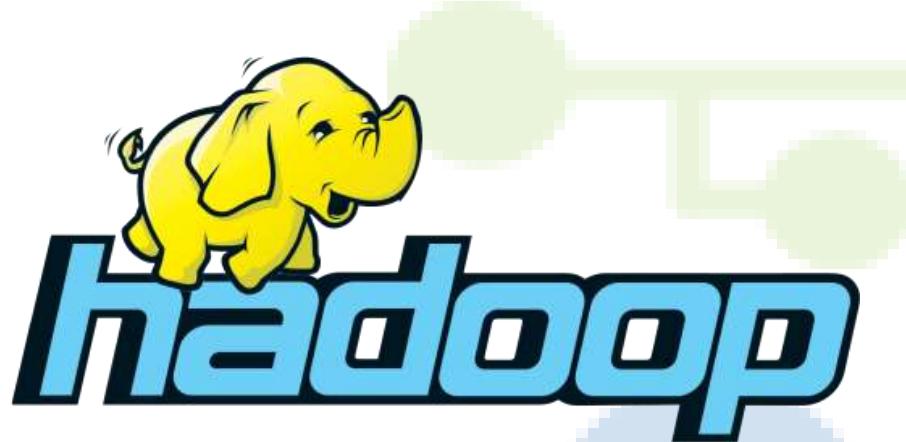
Introdução ao Hadoop



<http://hadoop.apache.org>

- Engenharia de Dados com Hadoop e Spark
- Formação Engenheiro de Dados

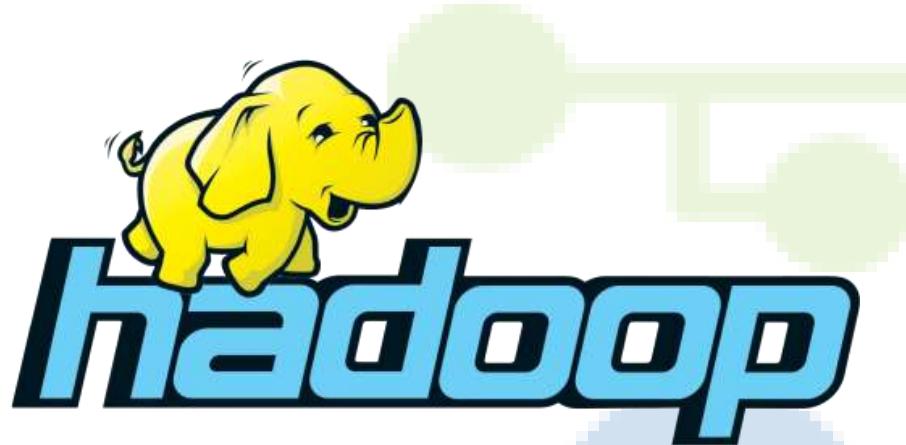
Introdução ao Hadoop



<http://hadoop.apache.org>

Apache Hadoop é um software open source para armazenamento e processamento em larga escala de grandes conjuntos de dados (Big Data), em clusters de hardware de baixo custo.

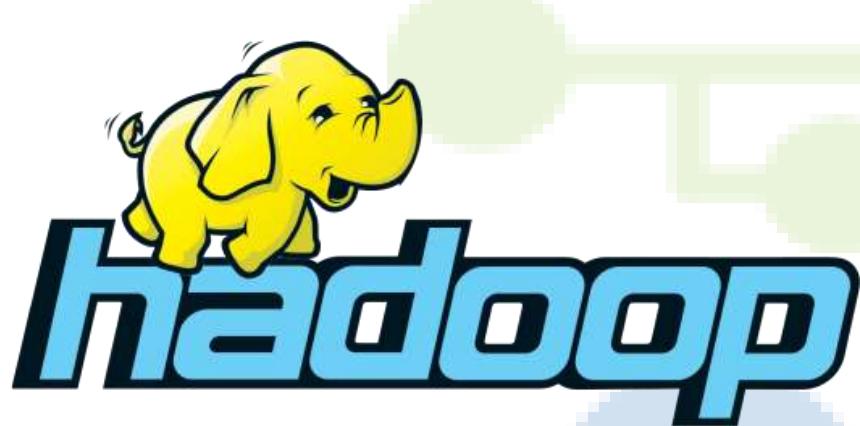
Introdução ao Hadoop



<http://hadoop.apache.org>

Temos visto o aumento crescente da capacidade de armazenamento dos discos rígidos.

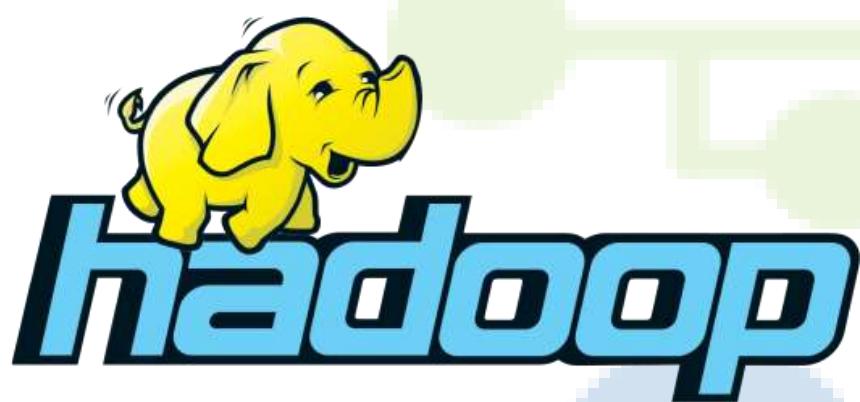
Introdução ao Hadoop



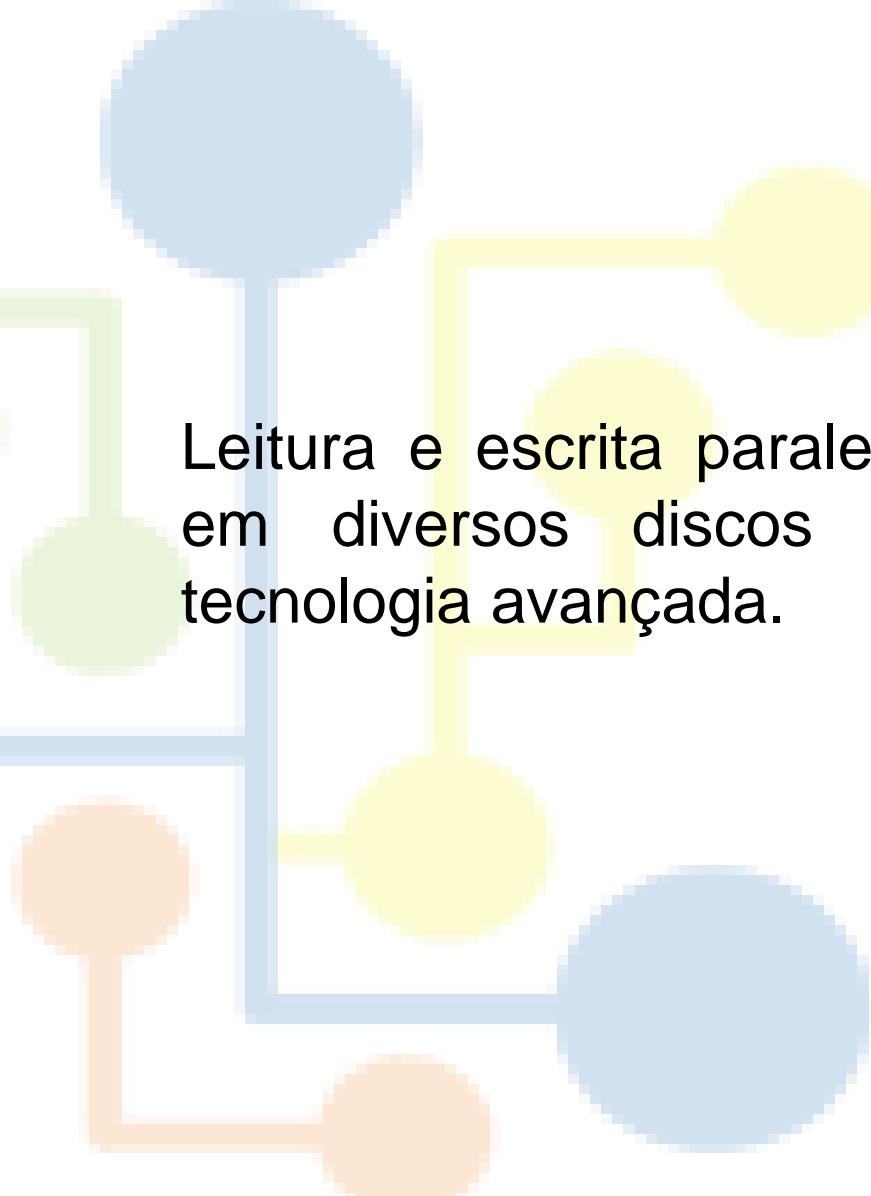
<http://hadoop.apache.org>

Mas a velocidade de leitura e escrita dos discos rígidos não tem crescido na mesma proporção.

Introdução ao Hadoop

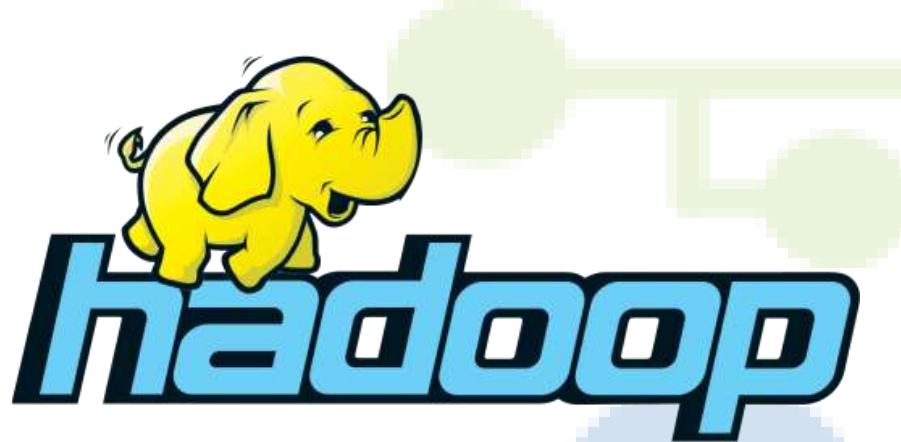


<http://hadoop.apache.org>

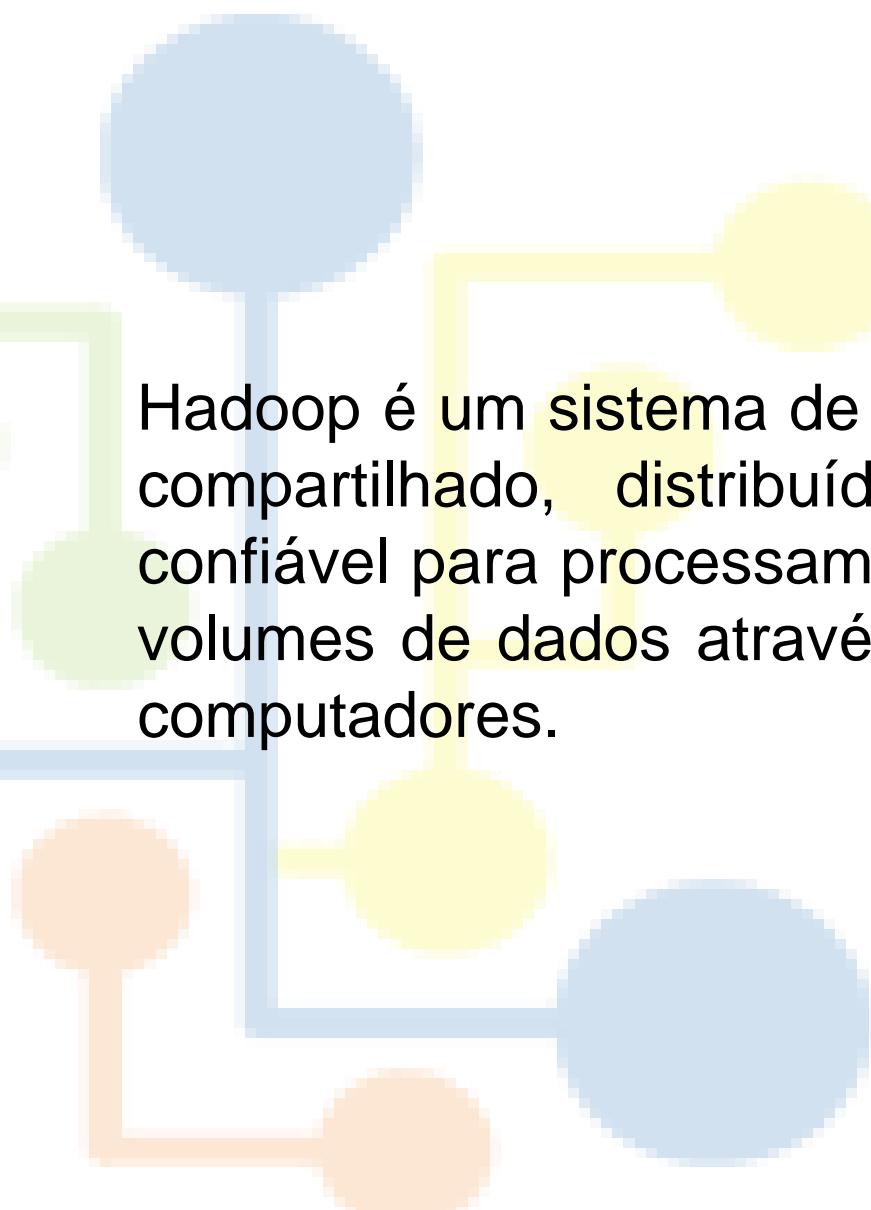
A diagram on the right side of the slide illustrates the concept of parallel and simultaneous access to multiple hard drives. It shows a central vertical blue line representing a data stream. This stream branches out into four horizontal lines, each ending in a colored circle (blue, green, yellow, and orange). These circles represent different hard drives. The diagram visually represents how Hadoop performs multiple read and write operations simultaneously across multiple storage devices.

Leitura e escrita paralela e simultânea
em diversos discos rígidos, requer
tecnologia avançada.

Introdução ao Hadoop



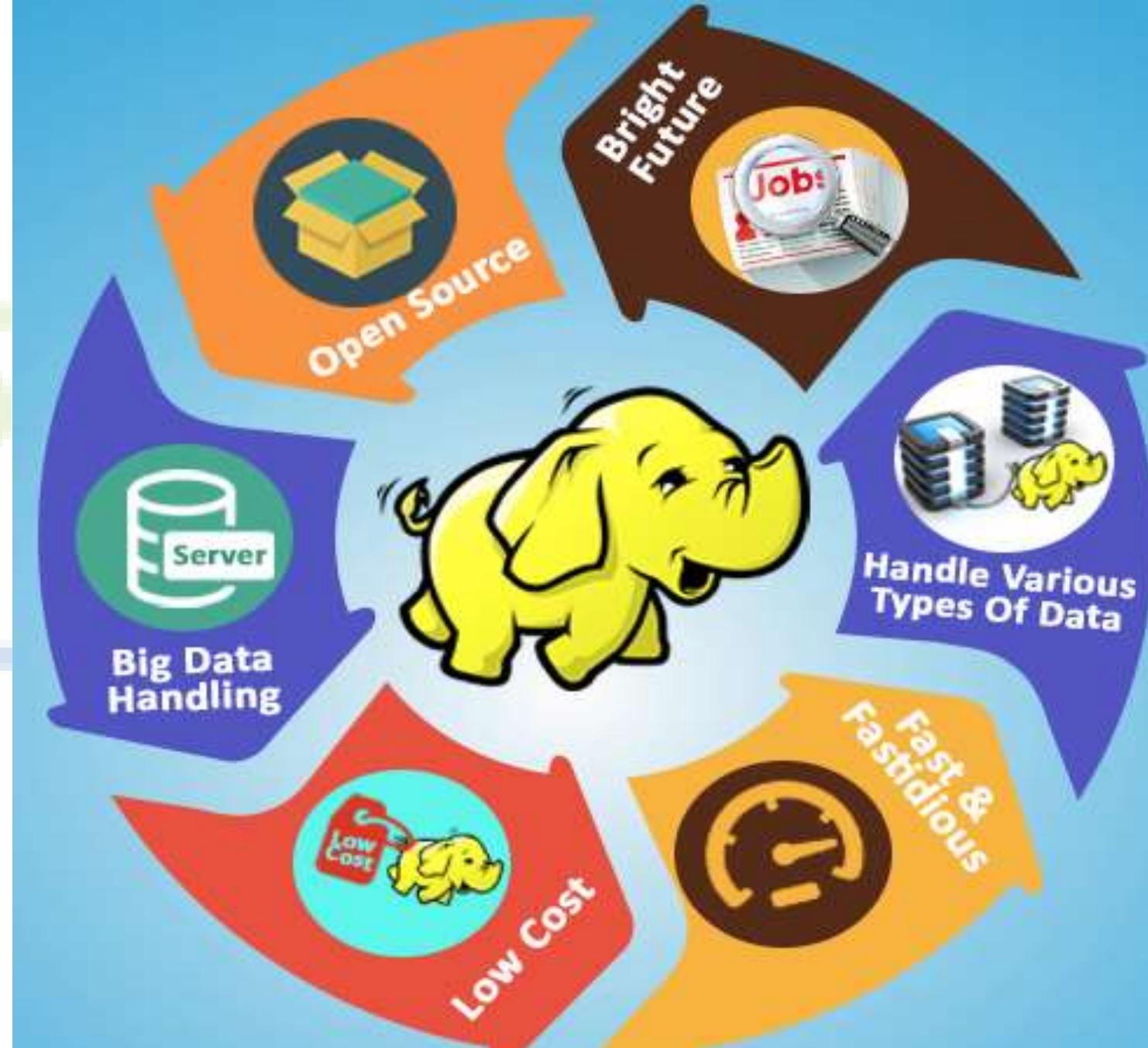
<http://hadoop.apache.org>

A diagram illustrating a distributed system architecture. It consists of several colored circles (blue, yellow, orange) representing nodes, connected by lines forming a network. A central blue circle is connected to a green square node, which is further connected to a yellow circle. This central structure is surrounded by other colored circles (blue, yellow, orange) and lines, representing a larger cluster of nodes.

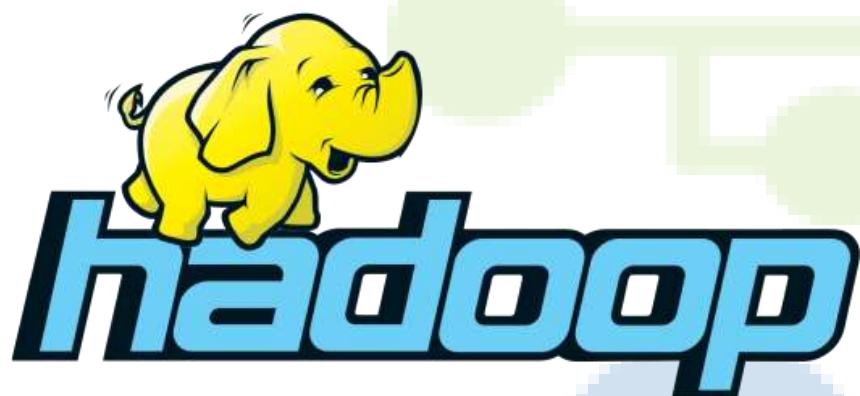
Hadoop é um sistema de armazenamento compartilhado, distribuído e altamente confiável para processamento de grandes volumes de dados através de clusters de computadores.

Introdução ao Hadoop

Em outras palavras, Hadoop é um framework que facilita o funcionamento de diversos computadores, com o objetivo de analisar grandes volumes de dados.



Introdução ao Hadoop



<http://hadoop.apache.org>

O projeto Apache hadoop é composto de 3 módulos principais:

- Hadoop Distributed File System (HDFS)
- Hadoop Yarn
- Hadoop MapReduce

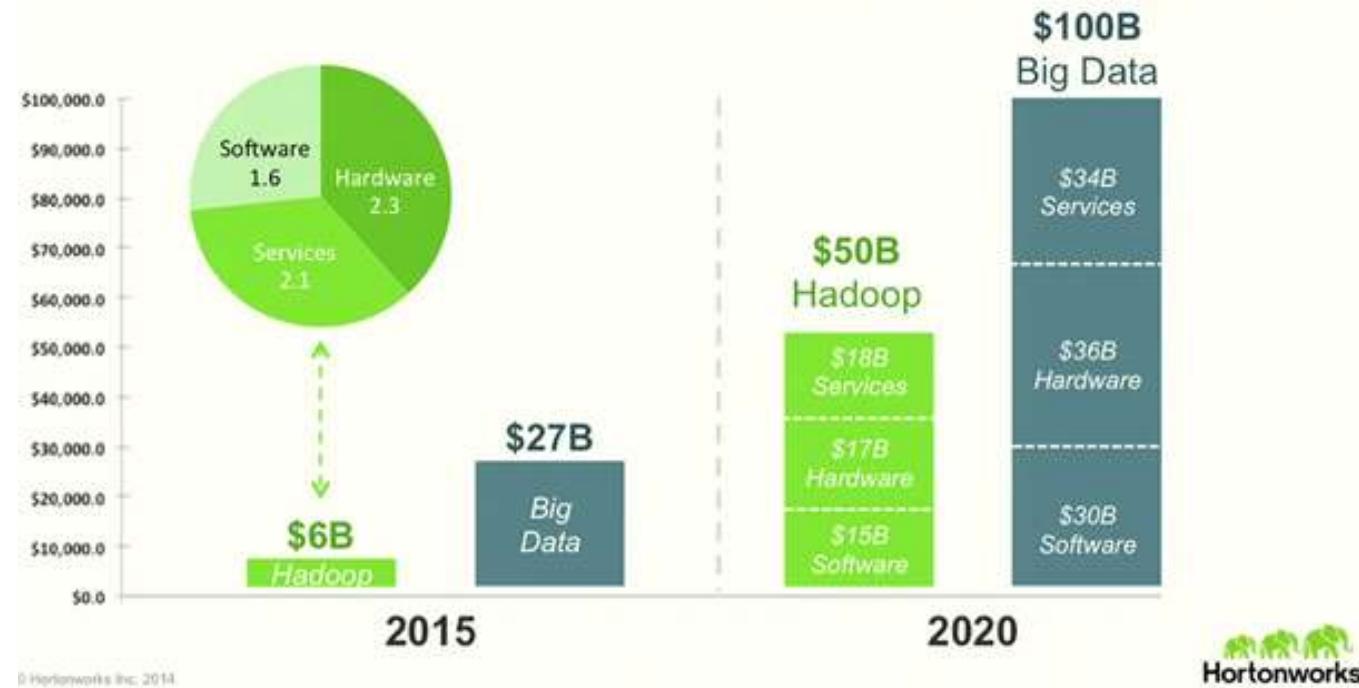
Introdução ao Hadoop

Hadoop is for problems too **Big** for traditional systems to handle

Introdução ao Hadoop

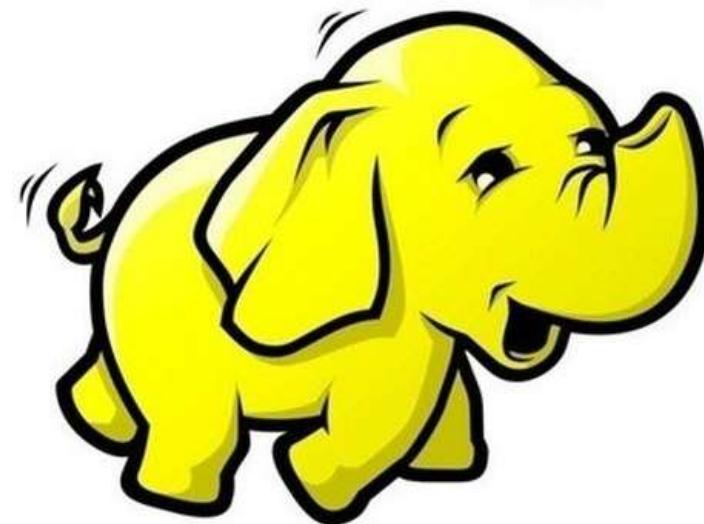
Pesquisas tem mostrado que o crescimento do Hadoop tem sido vertiginoso:

Big Data and Hadoop Markets Growing Sharply



Introdução ao Hadoop

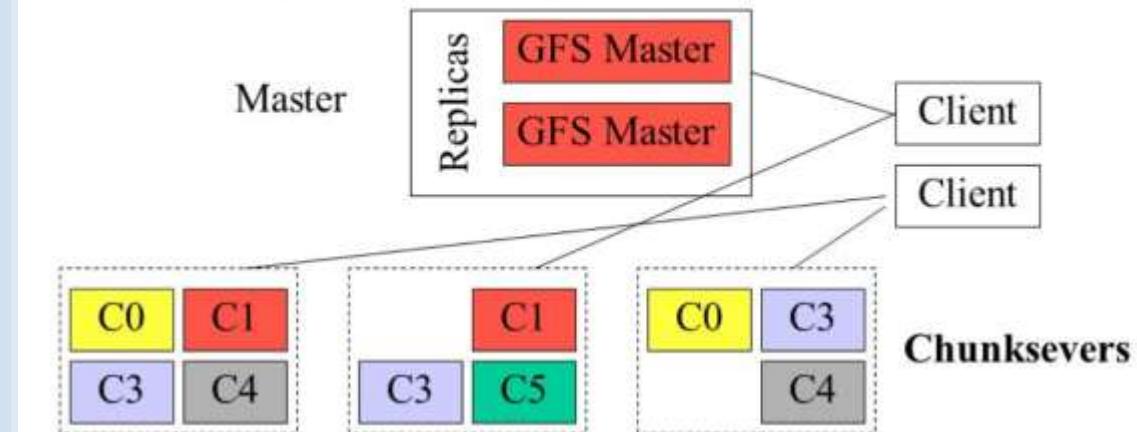
Hadoop é um framework gratuito, baseado em linguagem de programação Java, que suporta o processamento de grandes conjuntos de dados em ambientes de computação distribuída (através diversos computadores simultaneamente).

The logo for Hadoop, featuring the word "hadoop" in a bold, blue, lowercase sans-serif font.

Introdução ao Hadoop

Ele é baseado no Google File System (GFS)

Google File System (GFS)

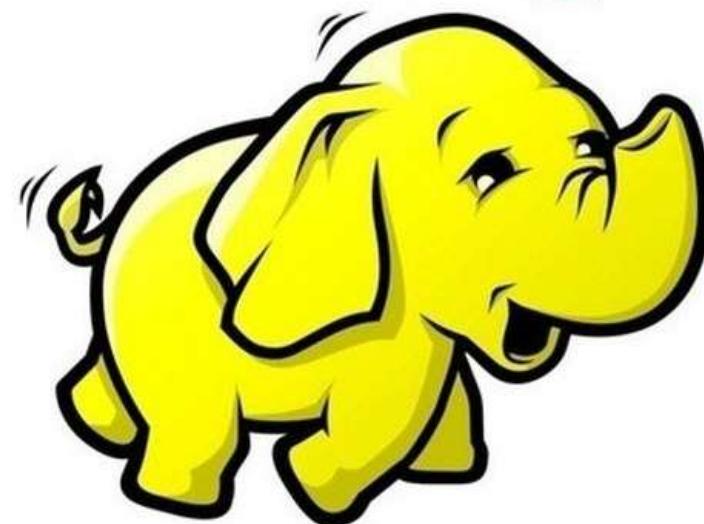


Introdução ao Hadoop

Hadoop permite executar aplicações em sistemas distribuídos através de diversos computadores (nodes), envolvendo petabytes de dados.

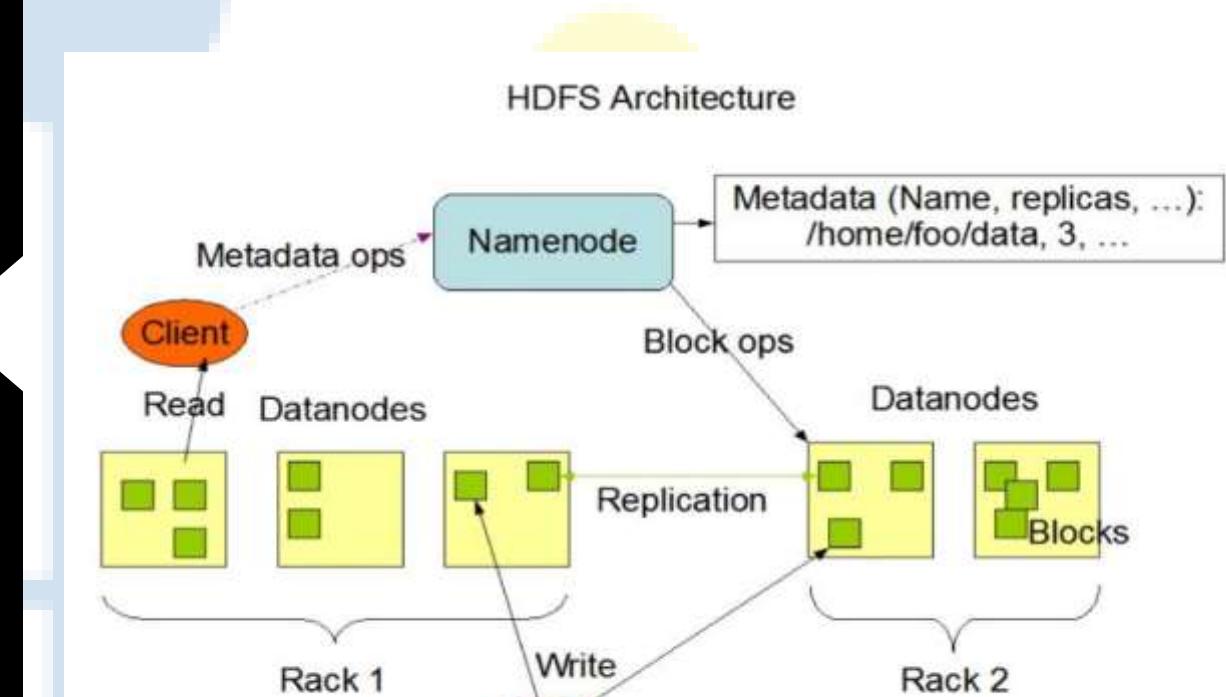
The Hadoop logo consists of the word "hadoop" in a bold, blue, lowercase sans-serif font. The letter 'h' has a black outline and a blue fill. The letters 'adoop' are also outlined in black and filled with blue.

hadoop



Introdução ao Hadoop

Hadoop utiliza o HDFS (Hadoop Distributed File System), que permite rápida transferência de dados entre os nodes. A segurança do Hadoop é feita com o Kerberos.

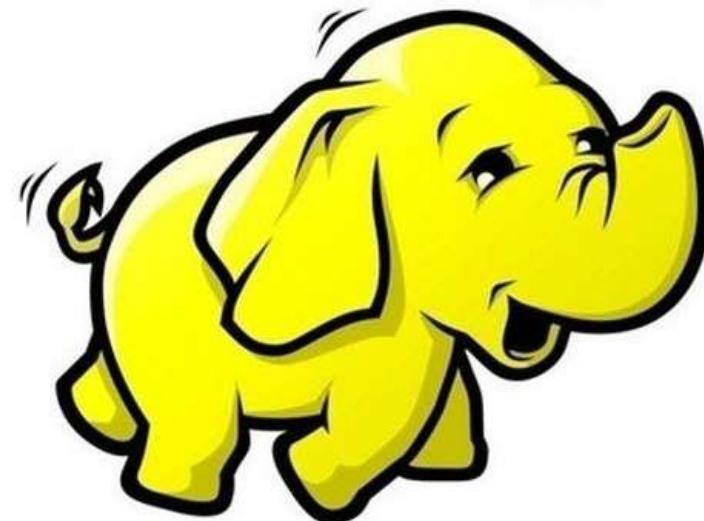


Introdução ao Hadoop

Hadoop é usado quando problemas muito grandes (Big) precisam de solução.

The Hadoop logo consists of the word "hadoop" in a bold, blue, lowercase sans-serif font. The letter "h" has a black outline and a white interior, while the other letters are solid blue with black outlines.

hadoop



Introdução ao Hadoop

Hadoop tem um baixo custo, não apenas por ser livre, mas por permitir o uso de hardware simples, computadores de baixo custo agrupados em cluster



Introdução ao Hadoop

Um das principais características do Hadoop é a confiabilidade e sua capacidade de se recuperar de falhas automaticamente.



Introdução ao Hadoop

Componentes Base do Hadoop:

Hadoop HDFS

Hadoop MapReduce

Introdução ao Hadoop

Componentes Base do Hadoop:



=

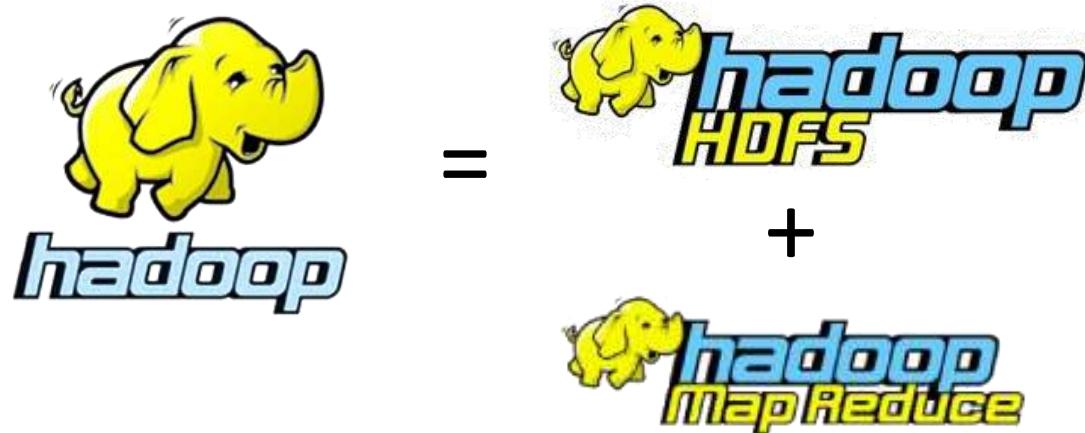


+



Introdução ao Hadoop

Componentes Base do Hadoop:



De forma bem simples, podemos dizer:

HDFS – armazenamento distribuído

MapReduce – computação distribuída

Introdução ao Hadoop

Por que o Hadoop está se tornando o padrão nos projetos de Big Data?

Baixo
Custo

Escalável

Tolerante a
Falhas

Flexível

Livre

Introdução ao Hadoop



- Tolerância a falhas e recuperação automática.
- Portabilidade entre hardware e sistemas operacionais heterogêneos.
- Escalabilidade para armazenar e processar grandes quantidades de dados.
- Confiabilidade, através da manutenção de várias cópias de dados.

Introdução ao Hadoop



- Flexibilidade – processa todos os dados independente do tipo e formato, seja estruturado ou não-estruturado.
- Confiabilidade - permite que os jobs sejam executados em paralelo e em caso de falhas de um job, outros não são afetados.
- Acessibilidade – suporte a diversas linguagens de programação como Java, C++, Python, Apache Pig.

Introdução ao Hadoop



Introdução ao Hadoop



HDFS (Hadoop Distributed File System)

- Foi desenvolvido utilizando o projeto do sistema de arquivos distribuídos (DFS). Ele é executado em hardware commodity (baixo custo). Ao contrário de outros sistemas distribuídos, HDFS é altamente tolerante a falha.

Introdução ao Hadoop



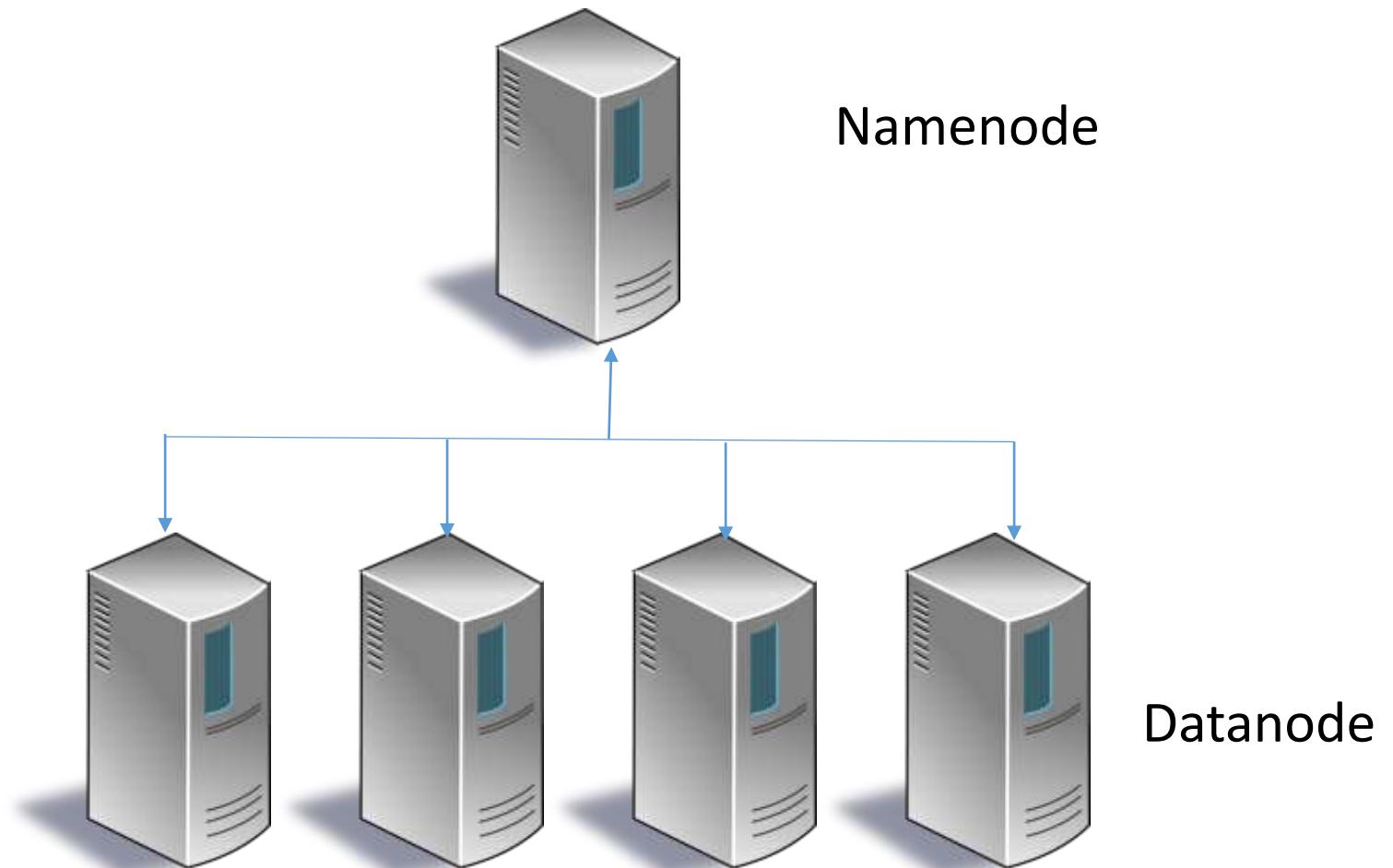
- **DFS (Distributed File System)** - foi criado para gestão de armazenamento em uma rede de computadores.
- **HDFS** é otimizado para armazenar grandes arquivos.
- **HDFS** foi pensado para executar em clusters de computadores de baixo custo.
- **HDFS** foi pensado para ser ótimo em performance do tipo WORM (Write Once, Read Many Times), que é um eficiente padrão de processamento de dados.
- **HDFS** foi pensando considerando o tempo de leitura de um conjunto de dados inteiro e não apenas o primeiro registro.

Introdução ao Hadoop

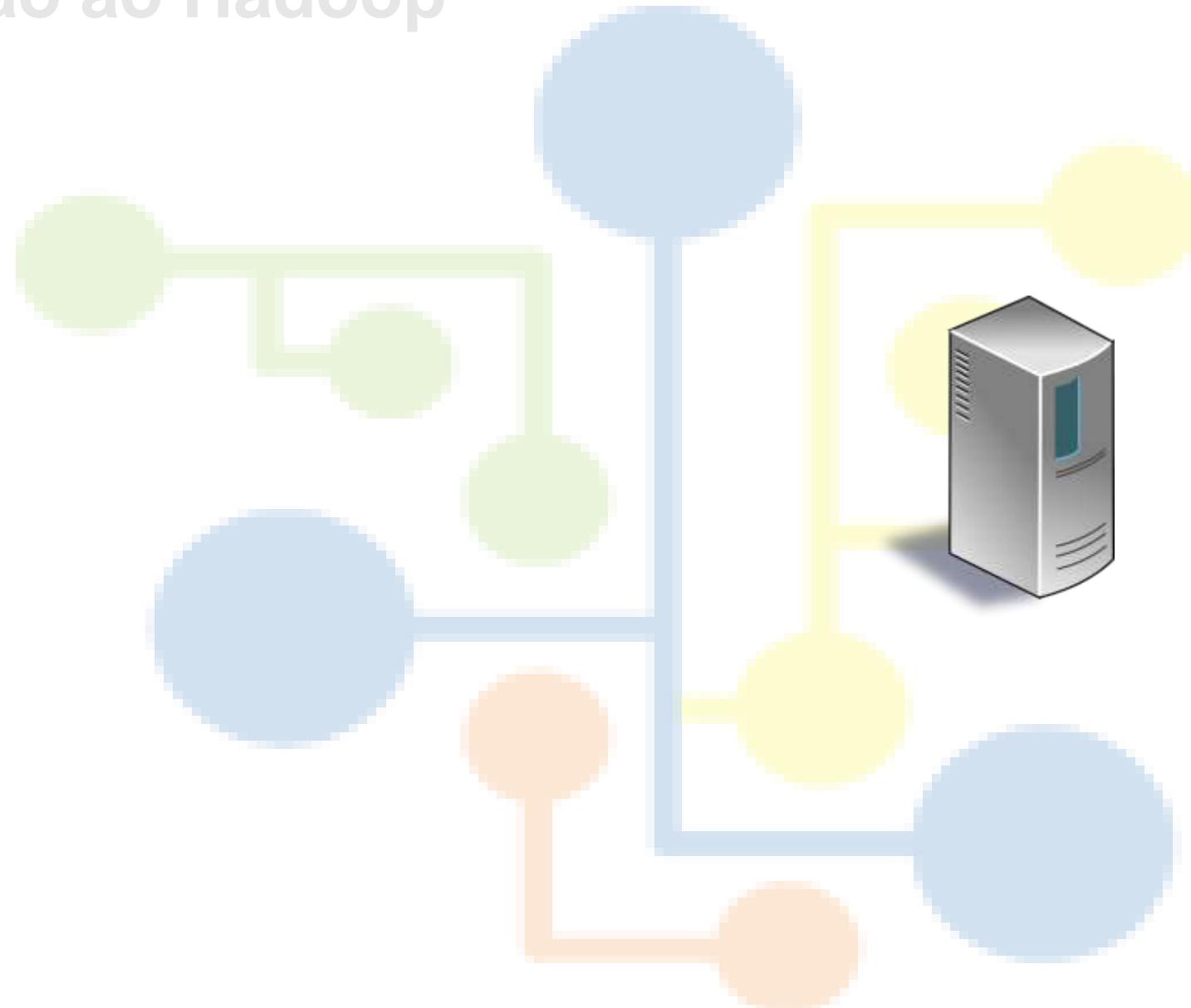


HDFS cluster possui 2 tipos de nodes:

Namenode (master node)
Datanode (worker node)

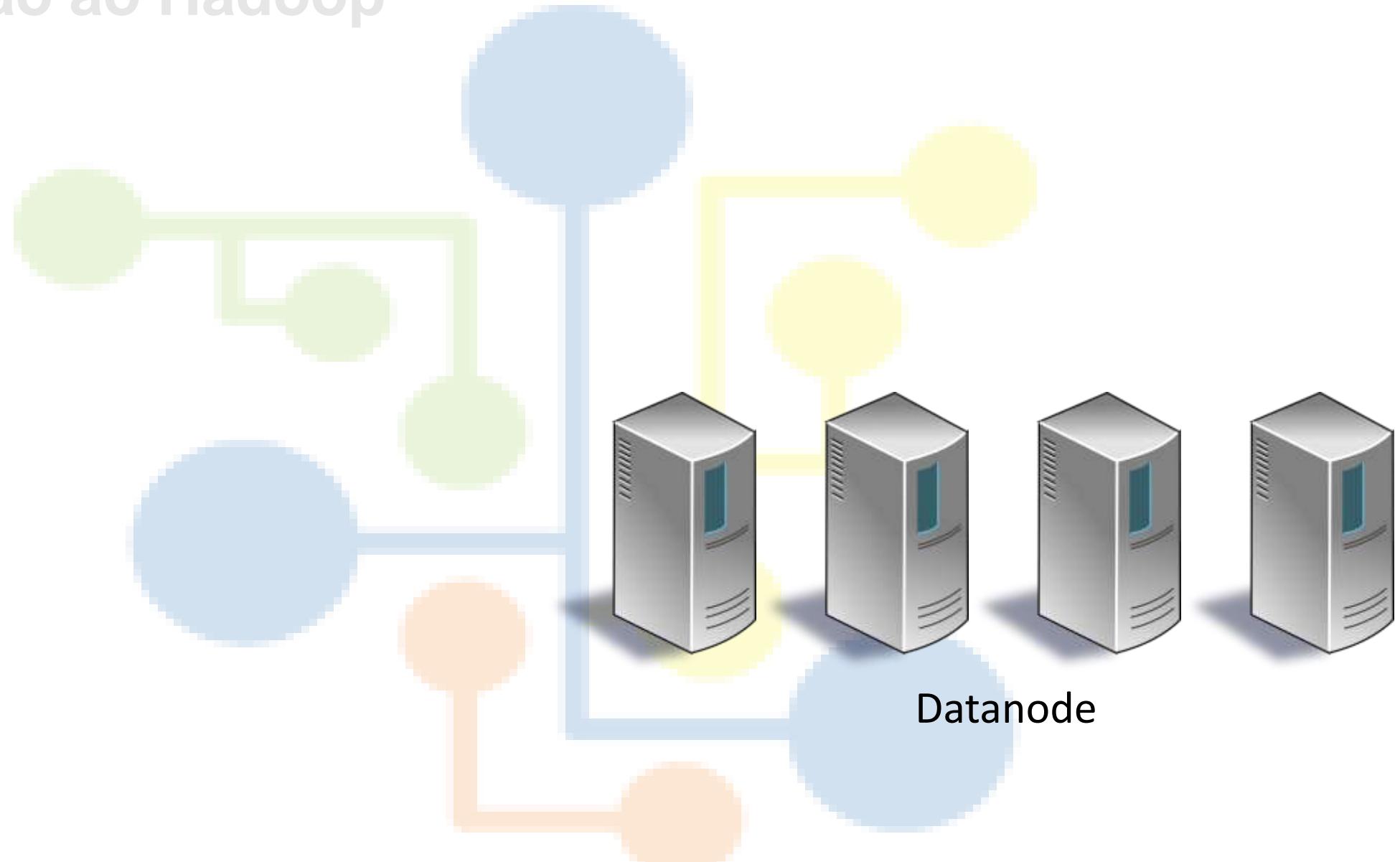


Introdução ao Hadoop



Namenode

Introdução ao Hadoop



Introdução ao Hadoop



Introdução ao Hadoop

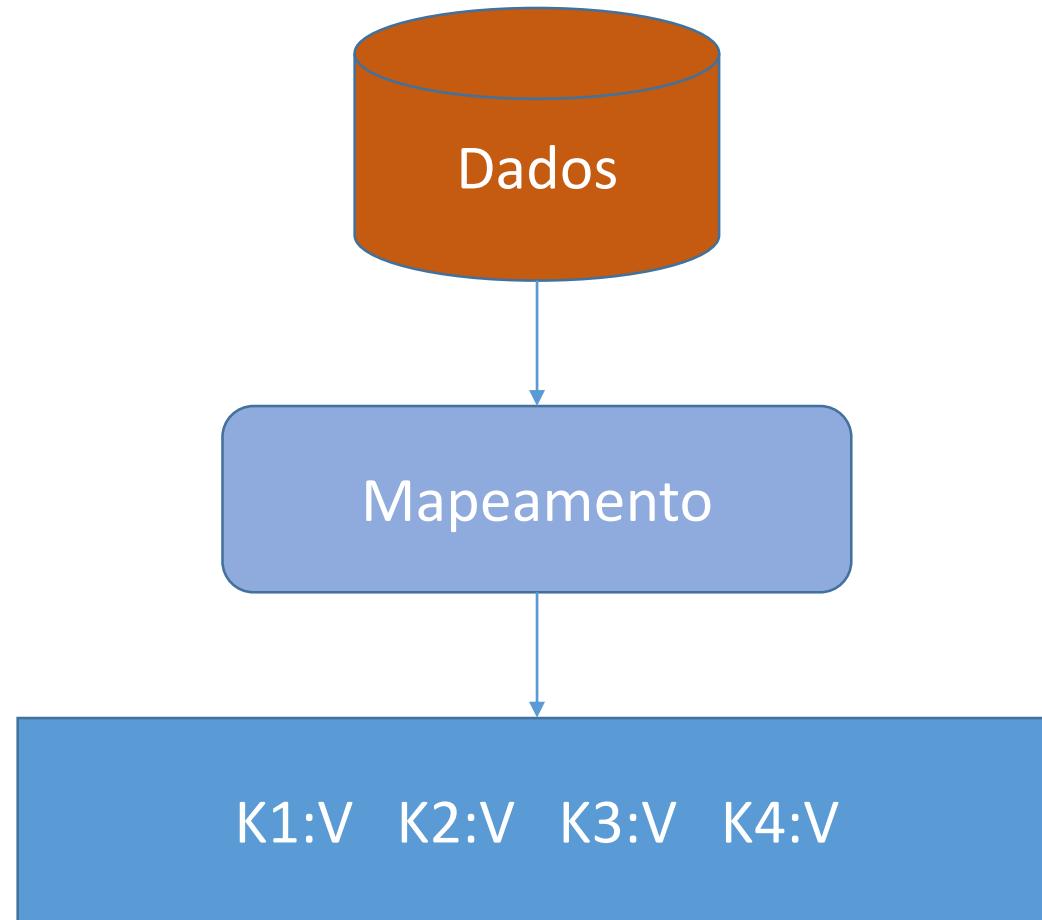


- MapReduce é um modelo de programação para processamento e geração de grandes conjuntos de dados.
- MapReduce transforma o problema de análise em um processo computacional que usa conjuntos de chaves e valores.
- MapReduce foi desenvolvido para tarefas que consomem minutos ou horas em computadores conectados em rede de alta velocidade gerenciados por um único master.
- MapReduce usa um tipo de análise de dados por força bruta. Todo o conjunto de dados é processado em cada query.
- Modelo de processamento em batch.

Introdução ao Hadoop

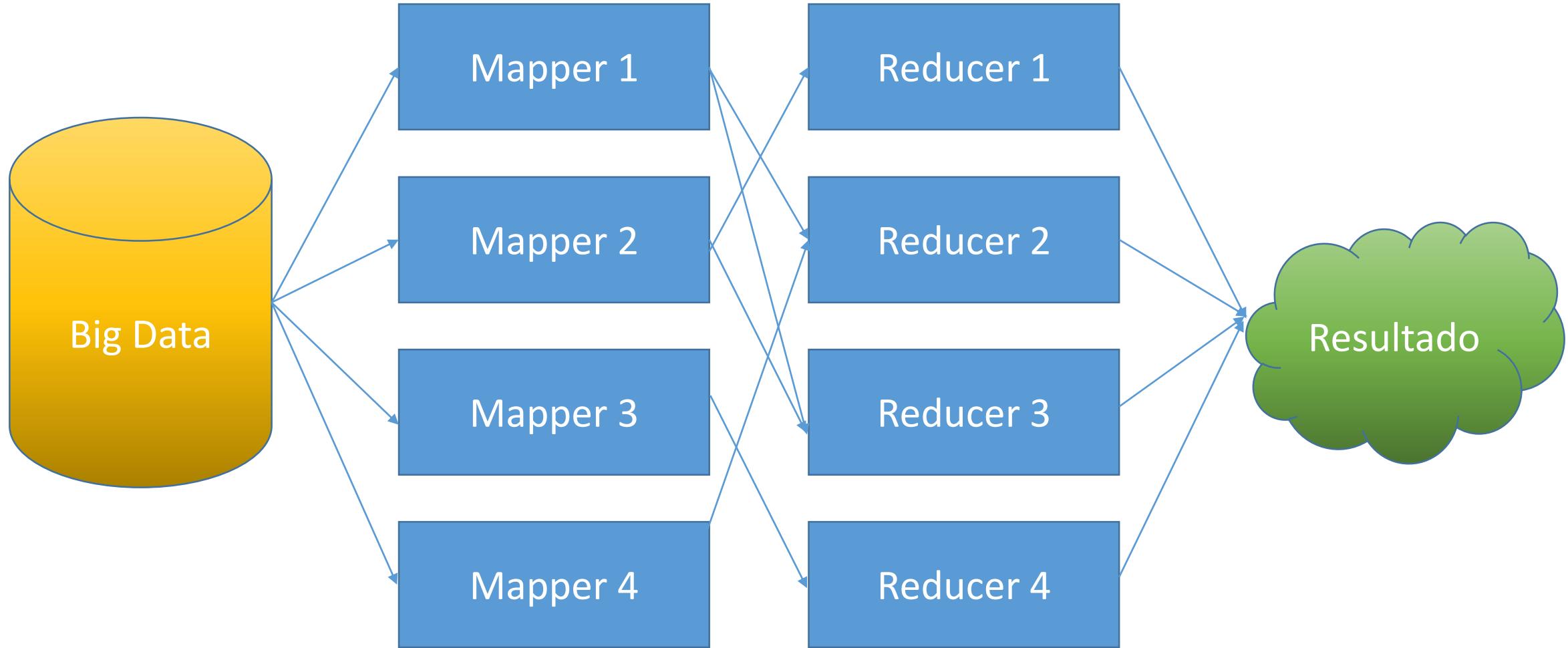


A função de
mapeamento, converte
dados em pares de
chave(K)/valor(V)



K = Key
V = Value

Introdução ao Hadoop



Seek Time x Transfer Rate

Introdução ao Hadoop



- MapReduce permite a execução de queries ad-hoc em todo o conjunto de dados em um tempo escalável.
- Muitos sistemas distribuídos combinam dados de múltiplas fontes (o que é bem complicado), mas MapReduce faz isso de forma eficiente e efetiva.
- O segredo da performance do MapReduce, está no balanceamento entre seeking e transfer: reduzir operações de seeking e usar de forma efetiva as operações de transfer.

Seek time – é o delay para encontrar um arquivo.

Transfer rate – é a velocidade para encontrar o arquivo.

Transfer rates tem melhorado显著mente
(é bem mais veloz que **Seek times**)

Introdução ao Hadoop



- O **MapReduce** é bom para atualizar todo (ou a maior parte) de um grande conjunto de dados.
- **RDBMS** (Relational Database Management System) são ótimos para atualizar pequenas porções de grandes bancos de dados.
- **RDBMS** utiliza o tradicional B-Tree, que é altamente dependente de operações de seek.
- **MapReduce** utiliza operações de SORT e Merge para recriar o banco de dados, o que é mais dependente de operações de transfer.

Introdução ao Hadoop



O MapReduce se baseia em operações de transfer,
o que deixa o acesso aos dados muito mais veloz.

MapReduce x RDBMS

| | RDBMS* | MapReduce |
|--------------------|----------------------------------|------------------------------------|
| Tamanho dos dados | Gigabytes (10^9) | Petabytes (10^{12}) |
| Acesso | Interativo e Batch | Batch |
| Updates | Leitura e Escrita diversas vezes | WORM (Write Once, Read Many Times) |
| Estrutura de Dados | Esquema estático | Esquema dinâmico |
| Integridade | Alta | Baixa |
| Escalabilidade | Não-linear | Linear |

* RDBMS = Relational Database Management System

Introdução ao Hadoop

Tipos de Dados

Dados Estruturados

Dados que são representados em formato tabular



Dados Semi Estruturados

Dados que não possuem um modelo formal de organização



Dados Não Estruturados

Dados sem estrutura pré-definida



Introdução ao Hadoop

MapReduce é muito efetivo com dados semi ou não estruturados!

Por que?



Introdução ao Hadoop

MapReduce interpreta dados durante as sessões de processamento de dados.

Ele não utiliza propriedades intrínsecas. Os parâmetros usados para selecionar os dados, são definidos pela pessoa que está fazendo a análise.

Hadoop não é um banco de dados.

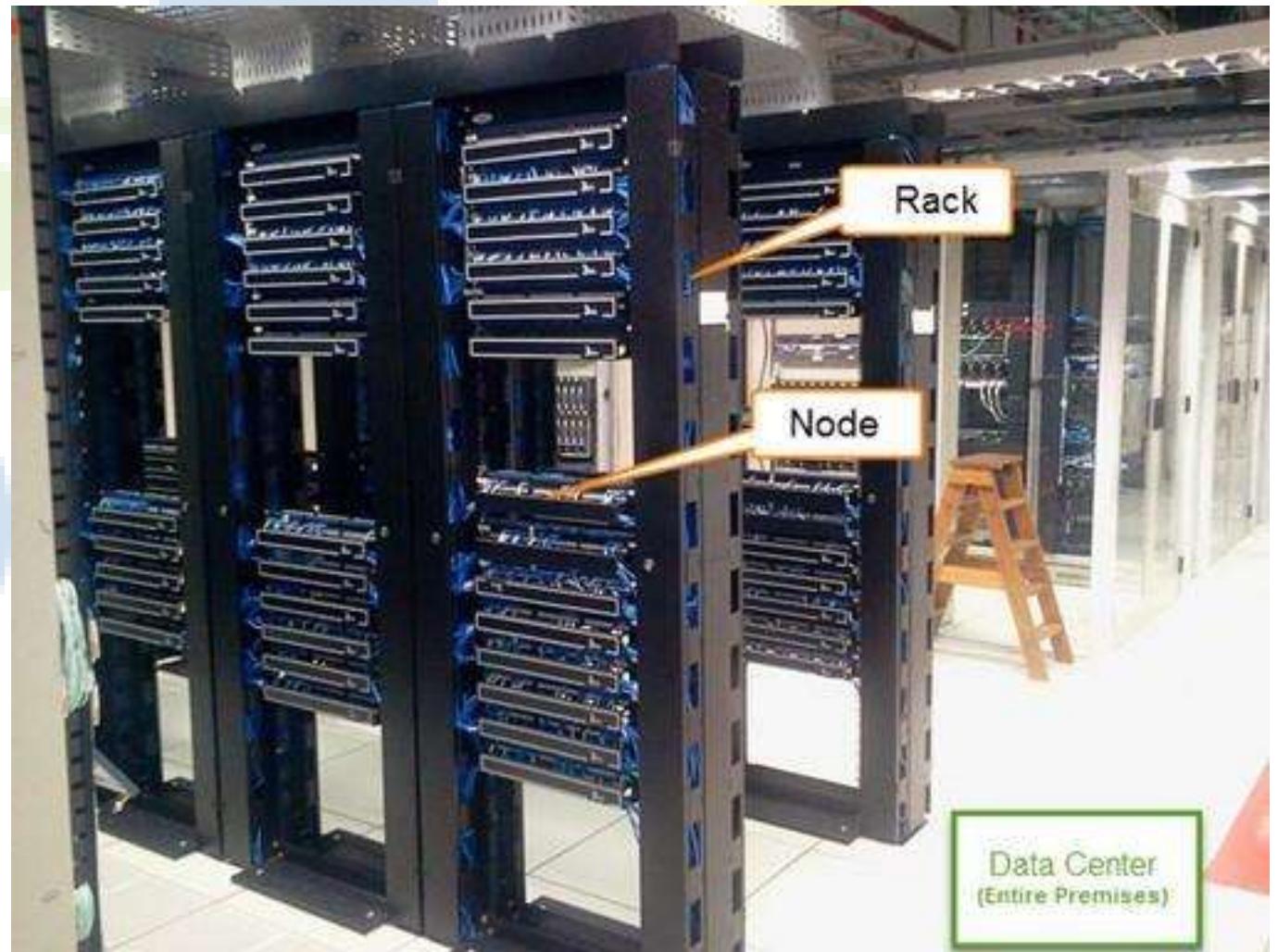
Hadoop é um framework para armazenamento e processamento de grandes conjuntos de dados!

Hadoop x RDBMS

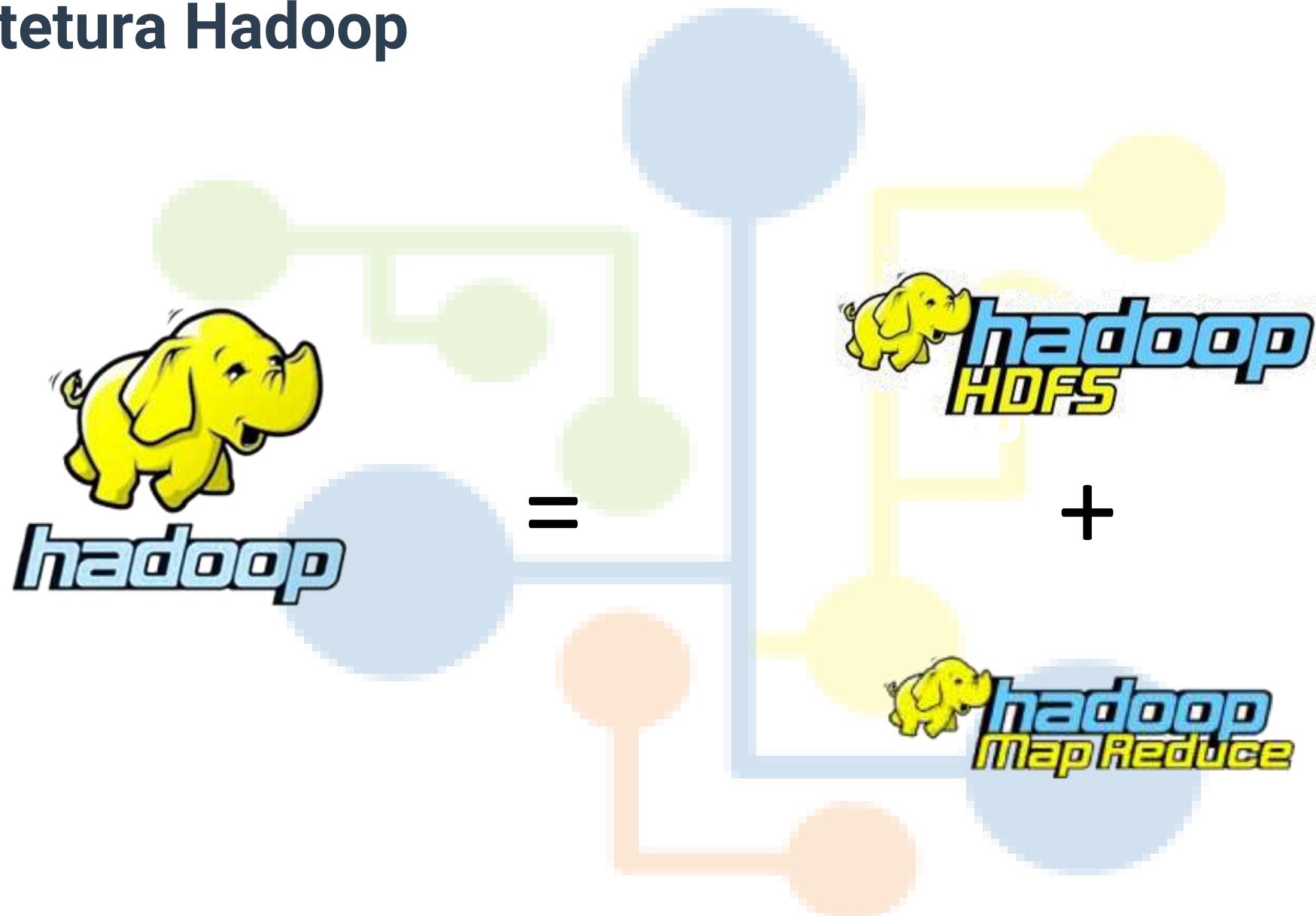
| | Hadoop | RDBMS |
|-----------------------------|---|---|
| Modelo de Computação | <ul style="list-style-type: none">▪ Conceito de Jobs▪ Cada Job é uma unidade de trabalho▪ Não há controle de concorrência | <ul style="list-style-type: none">▪ Conceito de transações▪ Uma transação é uma unidade de trabalho▪ Controle de concorrência |
| Modelo de Dados | <ul style="list-style-type: none">▪ Qualquer tipo de dado pode ser usado▪ Dados em qualquer formato▪ Modelo de apenas leitura | <ul style="list-style-type: none">▪ Dados estruturados com controle de esquema▪ Modelo de leitura/escrita |
| Modelo de Custo | <ul style="list-style-type: none">▪ Máquinas de custo mais baixo podem ser usadas | <ul style="list-style-type: none">▪ Servidores de maior custo são necessários |
| Tolerância a Falhas | <ul style="list-style-type: none">▪ Simples, mas eficiente mecanismo de tolerância a falha | <ul style="list-style-type: none">▪ Falhas são raras de ocorrer▪ Mecanismos de recuperação |

Arquitetura Hadoop

Mas o que é um Cluster afinal?



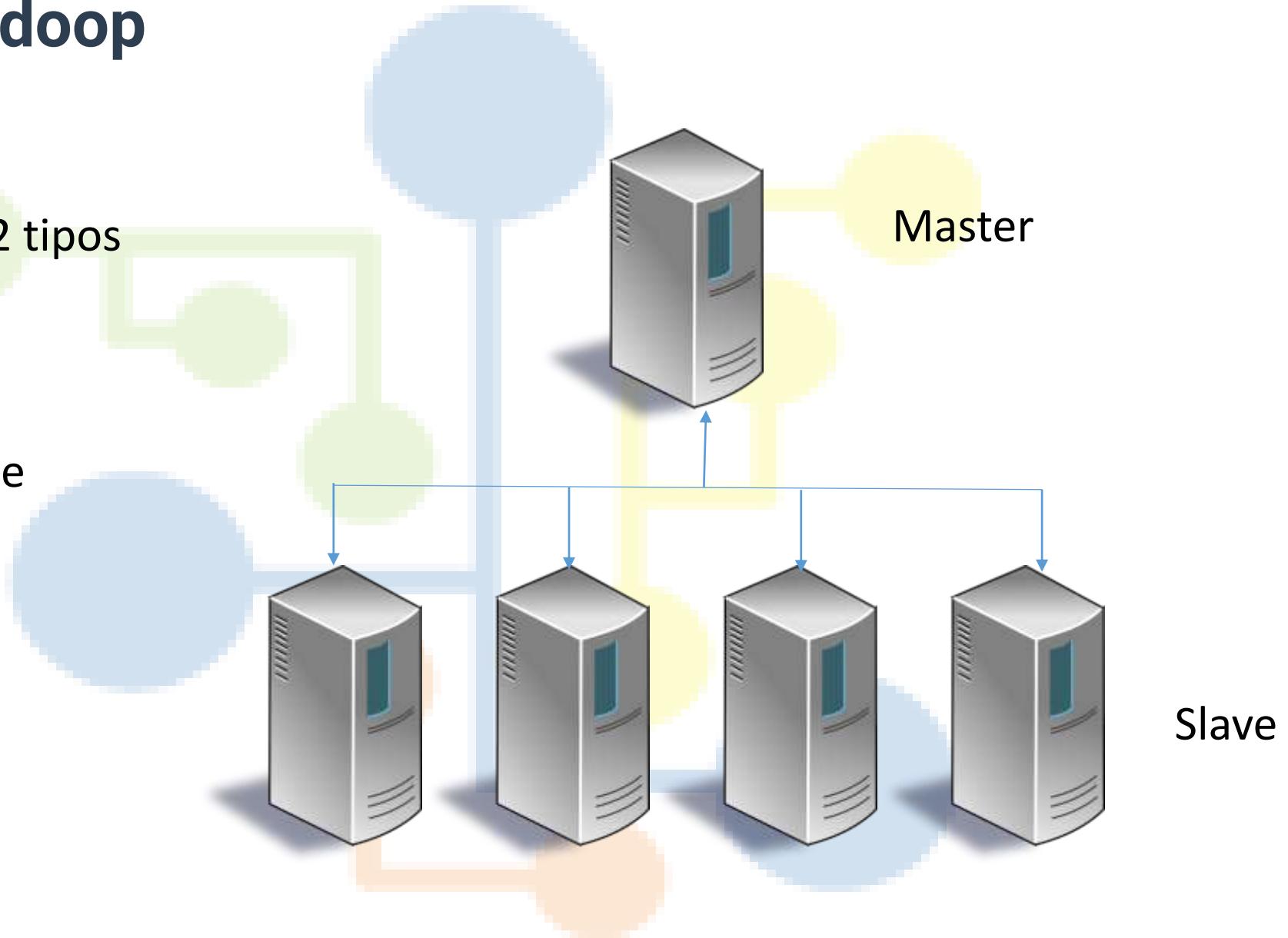
Arquitetura Hadoop



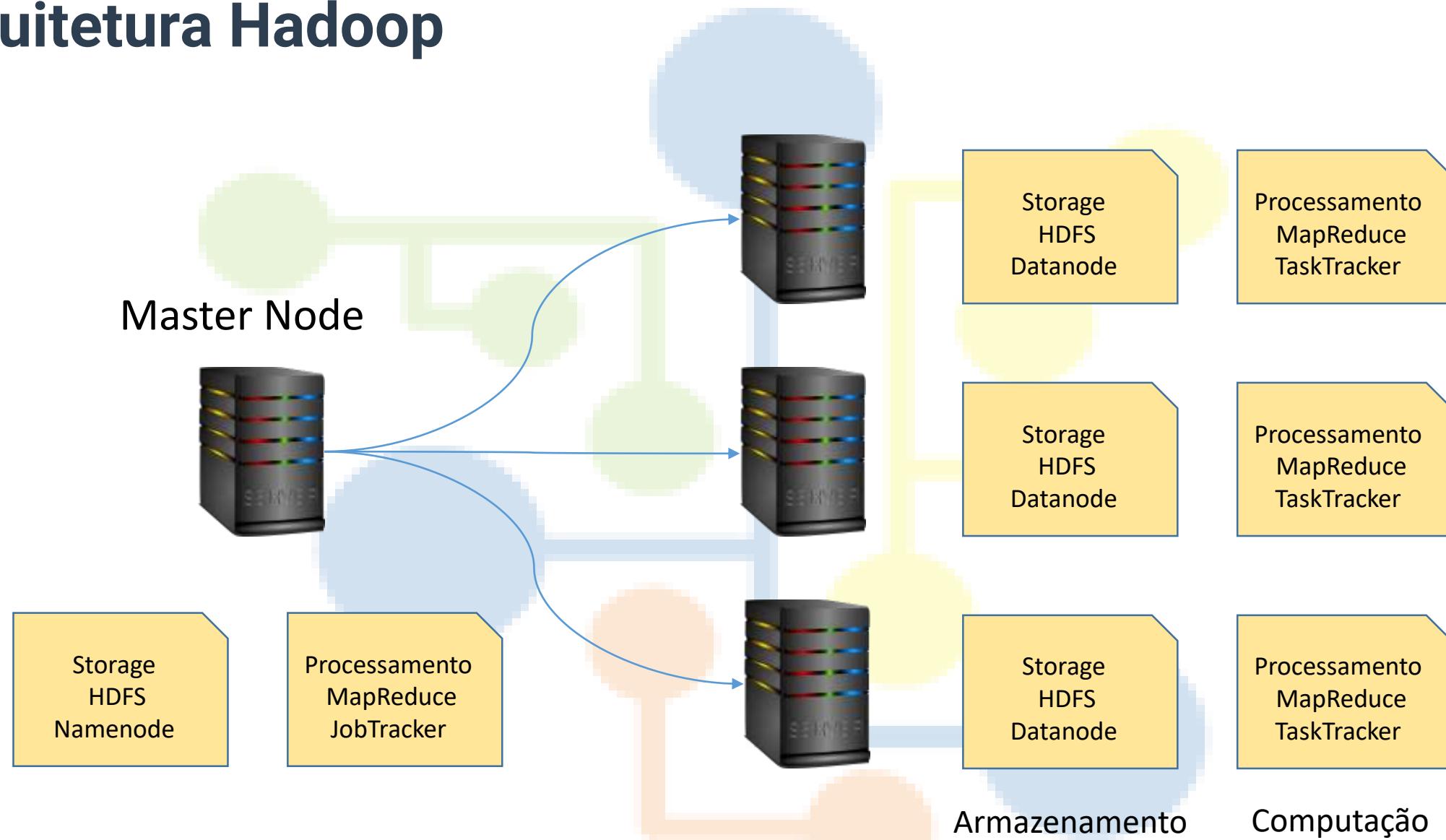
Arquitetura Hadoop

Cluster Hadoop possui 2 tipos de nodes:

Master node
Worker (slave) node

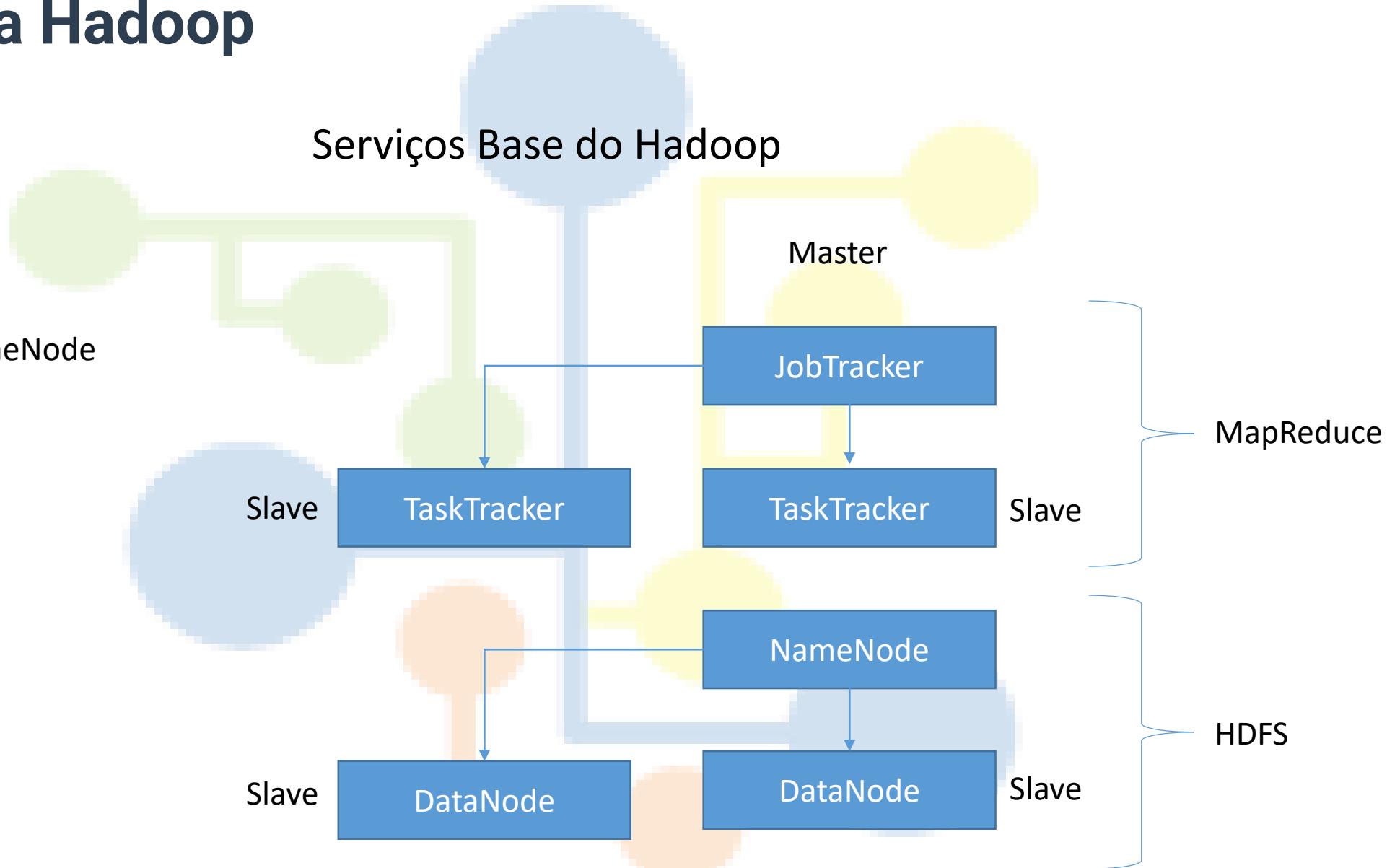


Arquitetura Hadoop

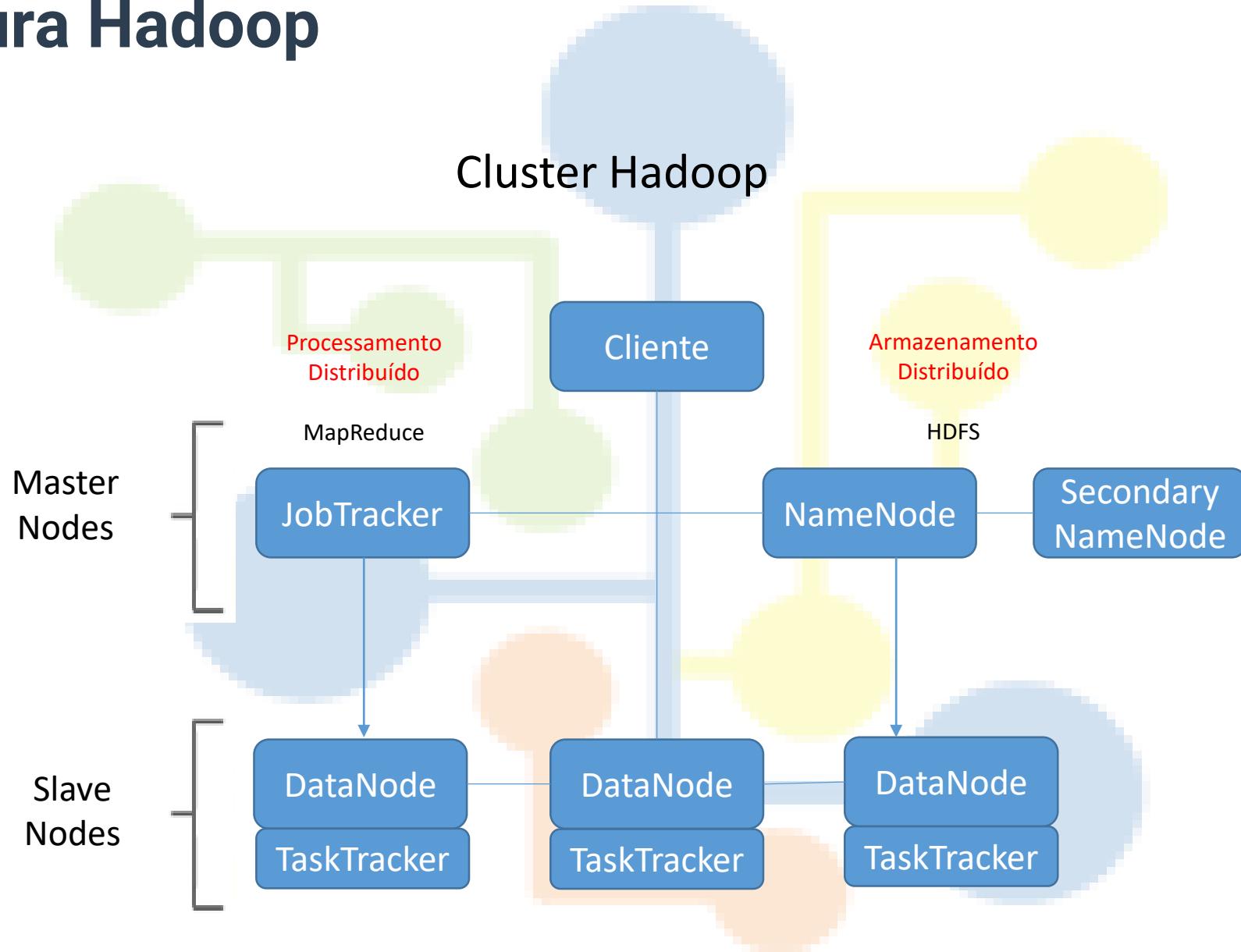


Arquitetura Hadoop

- NameNode
- Secondary NameNode
- DataNode
- JobTracker
- TaskTracker

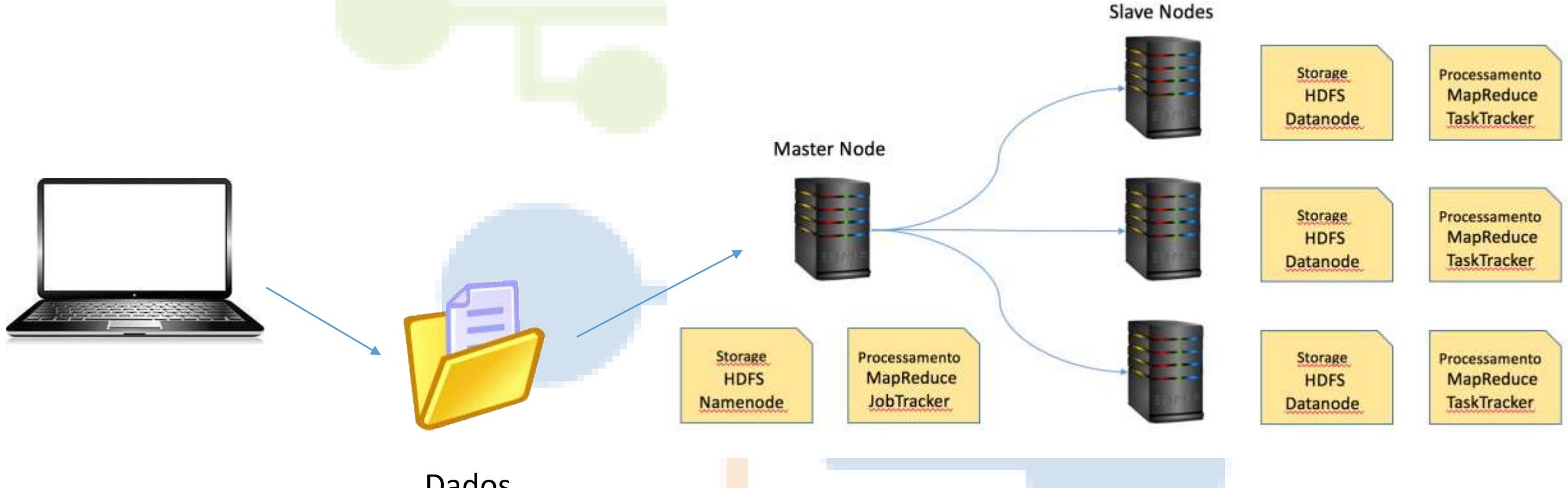


Arquitetura Hadoop

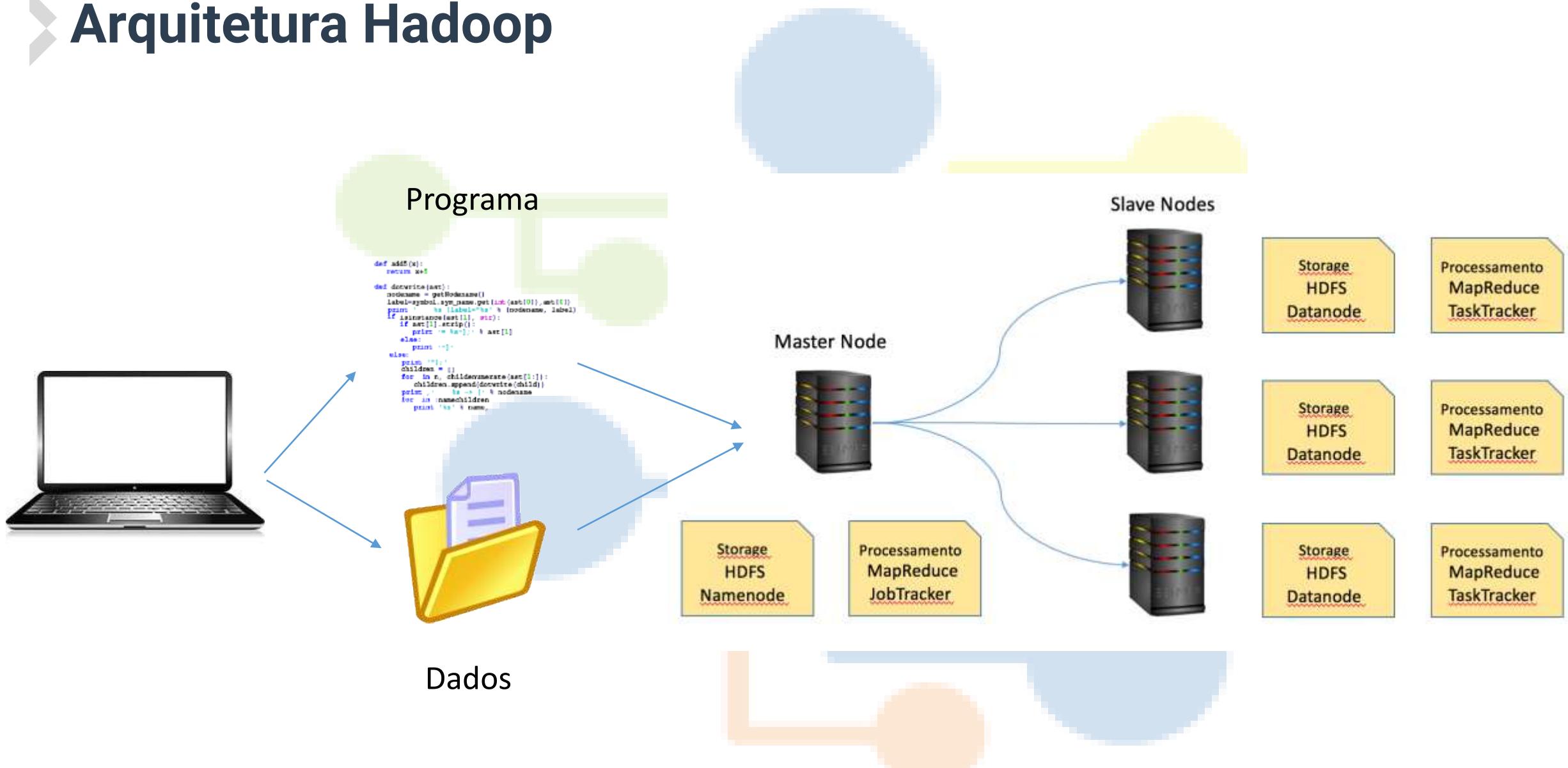


Arquitetura Hadoop

Passo 1 – Dados são enviados para o cluster Hadoop



Arquitetura Hadoop



Arquitetura Hadoop

Modos de Configuração do Hadoop

Hadoop suporta 3 modos de configuração:

Modo Standalone

Todos os serviços Hadoop são executados em uma única JVM, no mesmo servidor

Pseudo Distribuído

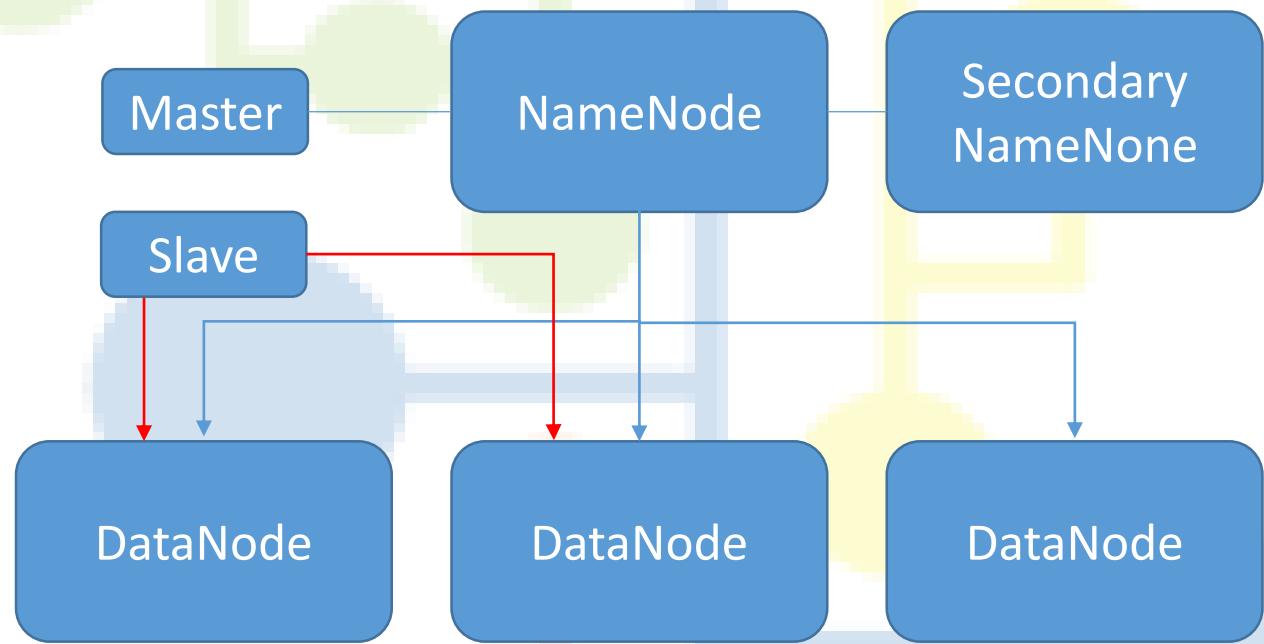
Serviços individuais do Hadoop são atribuídos a JVM's individuais, no mesmo servidor

Totalmente Distribuído

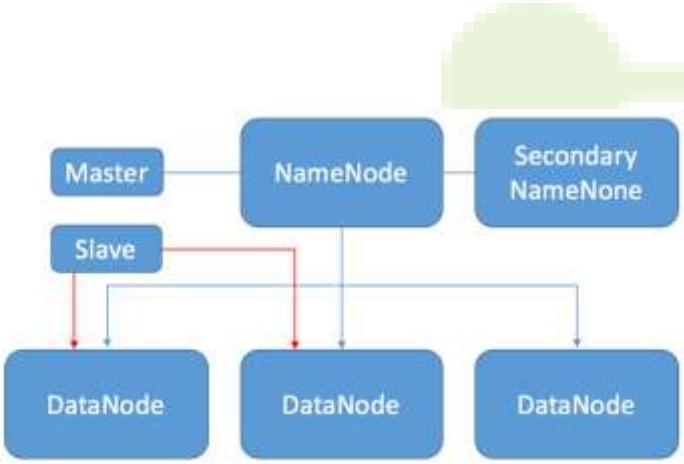
Serviços individuais do Hadoop são executados em JVM's individuais, mas através de cluster

Arquitetura Hadoop

Arquitetura HDFS

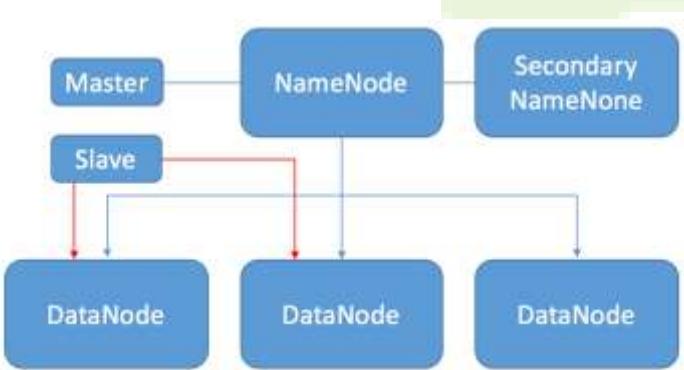


Arquitetura Hadoop



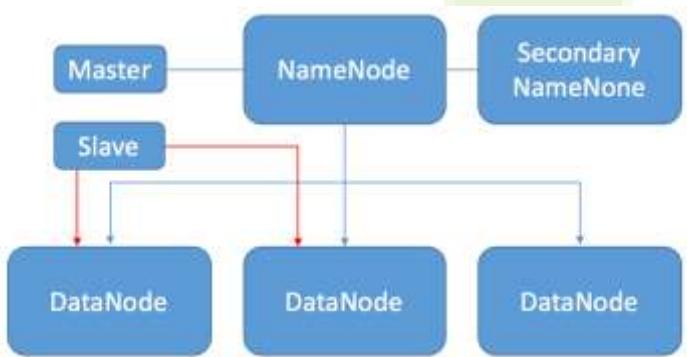
1. Os serviços NameNode e Secondary NameNode, constituem os serviços Master. Os serviços DataNode são os slaves.

Arquitetura Hadoop



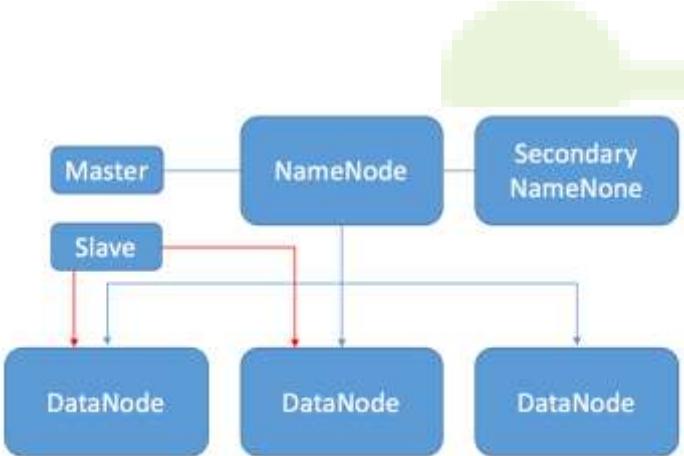
2. O serviço Master é responsável por aceitar os Jobs das aplicações clientes e garantir que os dados requeridos para a operação sejam carregados e segregados em pedaços de blocos de dados.

Arquitetura Hadoop



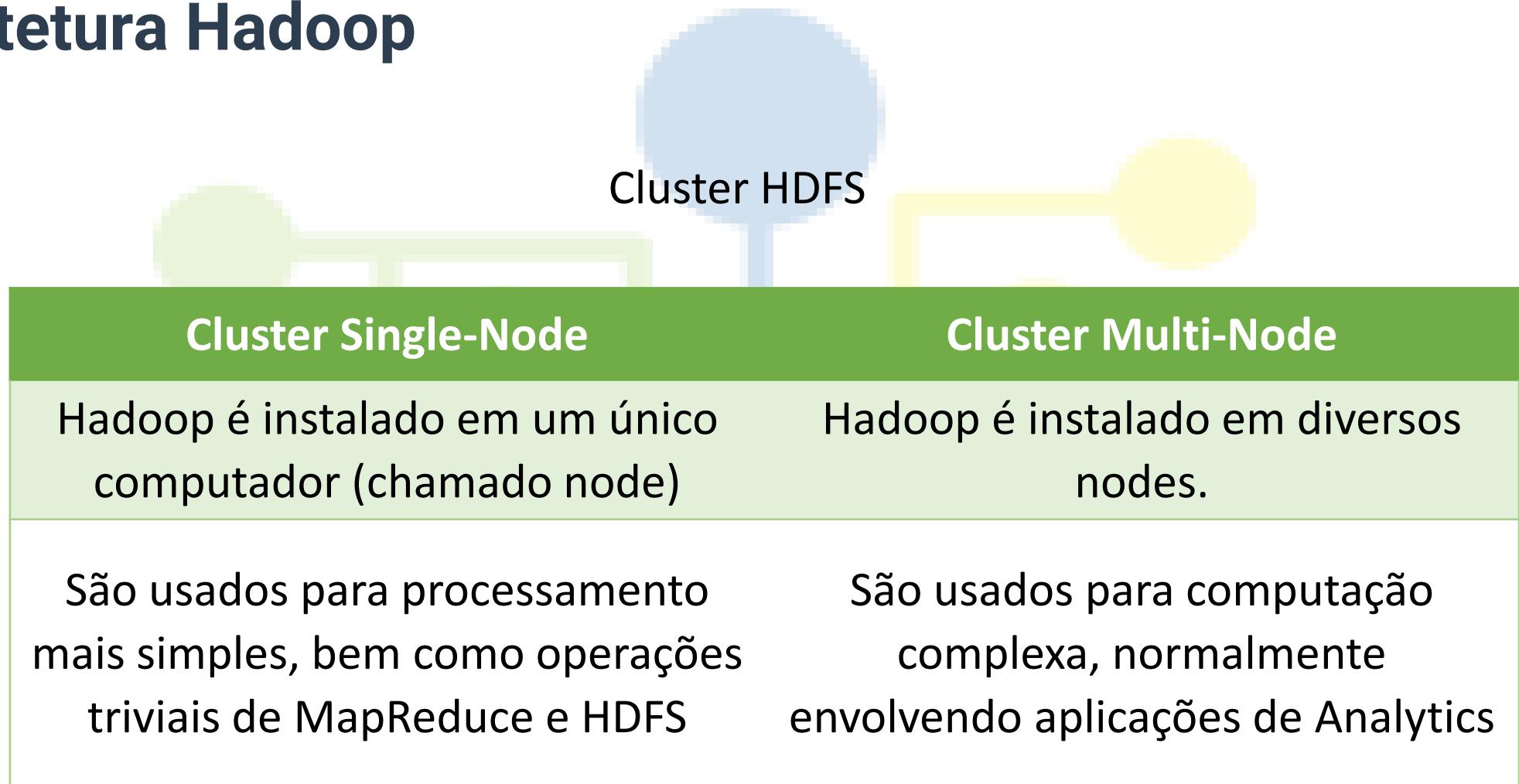
3. O HDFS permite que os dados sejam armazenados em arquivos. Um arquivo é dividido em um ou mais blocos que são armazenados e replicados pelos DataNodes. Os blocos de dados são então distribuídos para o sistema de DataNodes dentro do cluster. Isso garante que as réplicas de dados sejam mantidas.

Arquitetura Hadoop



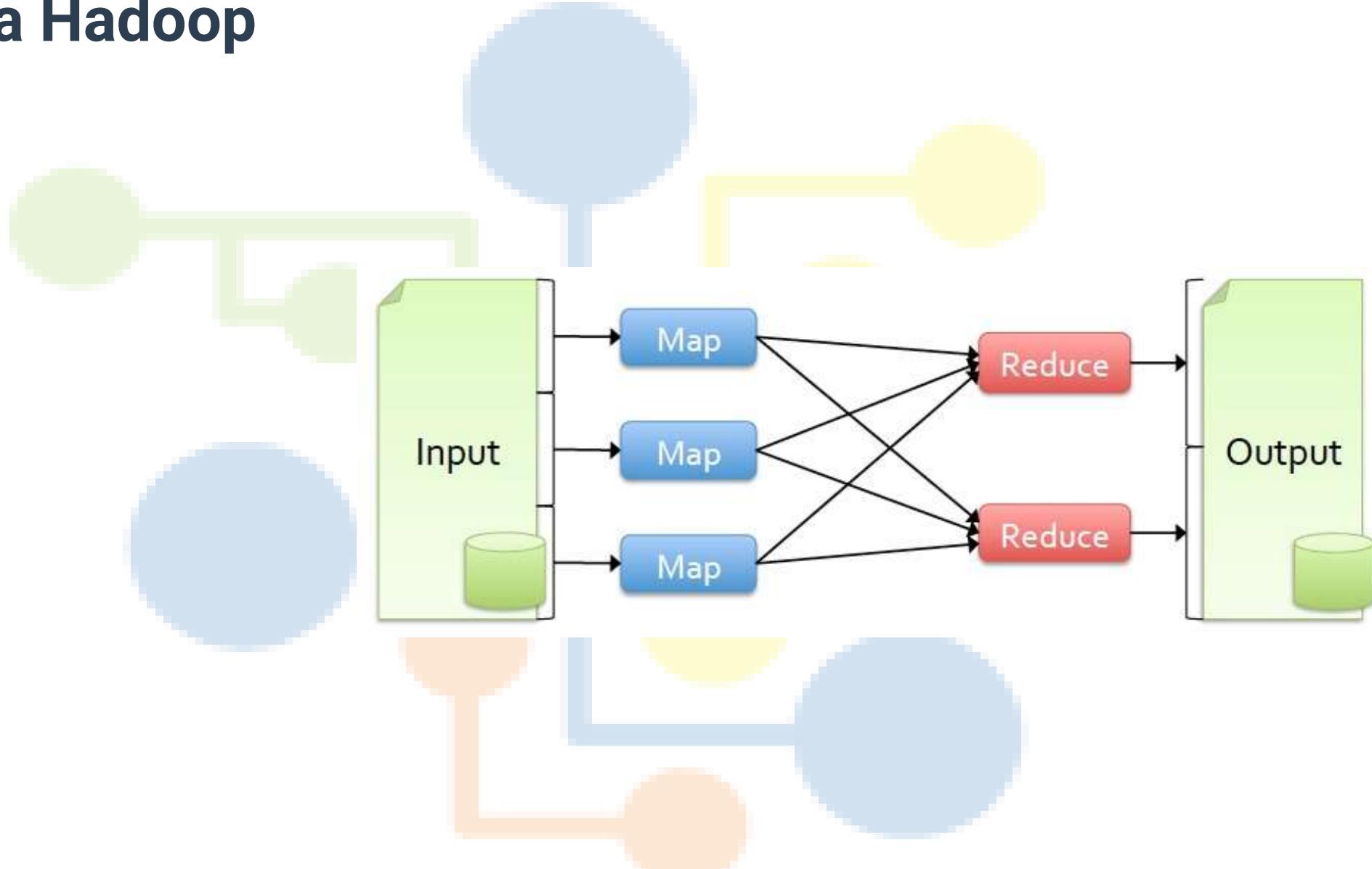
-
4. As réplicas de cada bloco de dados são distribuídas em computadores em todo o cluster para permitir o acesso de dados confiável e de forma rápida.

Arquitetura Hadoop

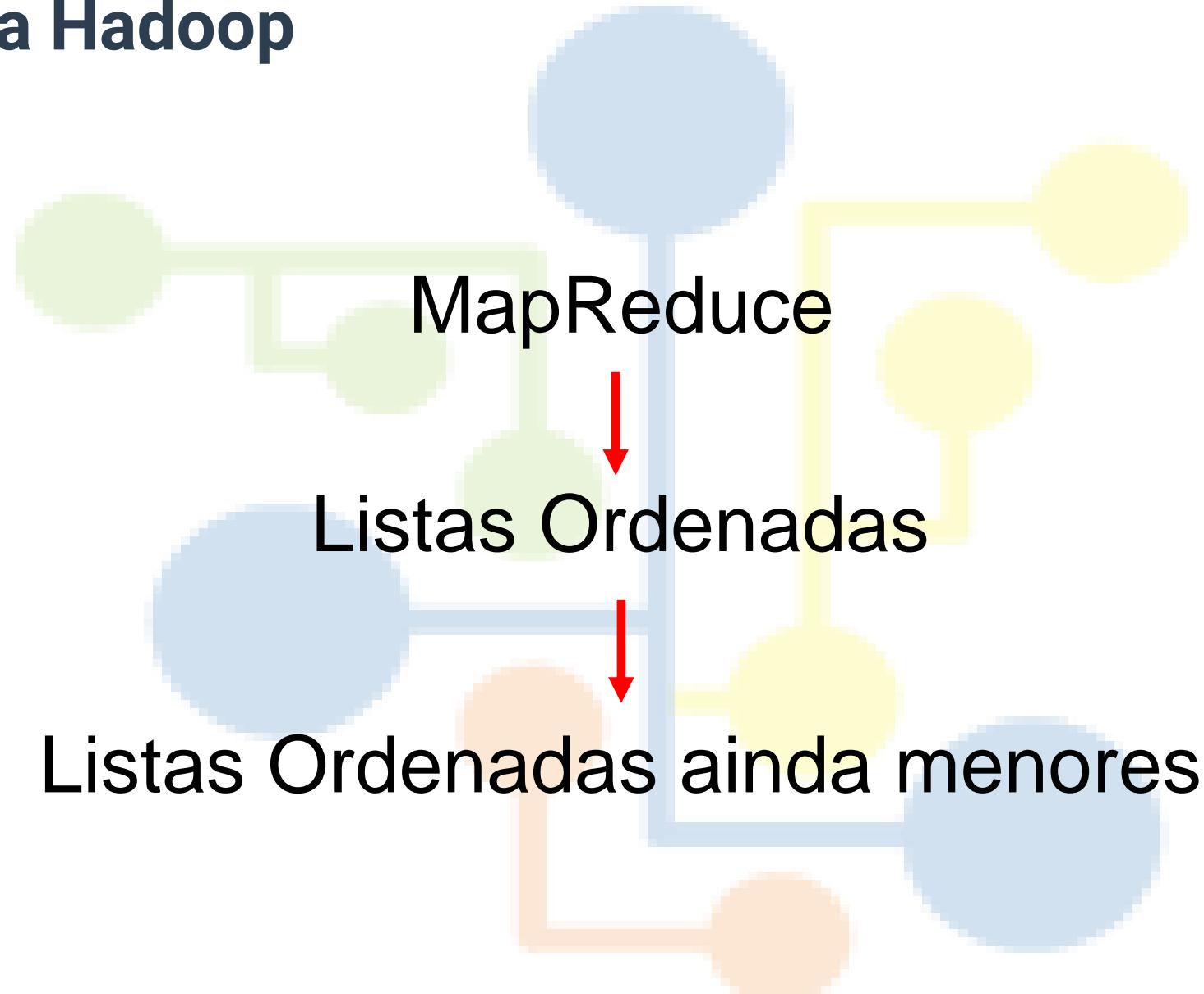


| Cluster Single-Node | Cluster Multi-Node |
|---|---|
| Hadoop é instalado em um único computador (chamado node) | Hadoop é instalado em diversos nodes. |
| São usados para processamento mais simples, bem como operações triviais de MapReduce e HDFS | São usados para computação complexa, normalmente envolvendo aplicações de Analytics |

Arquitetura Hadoop

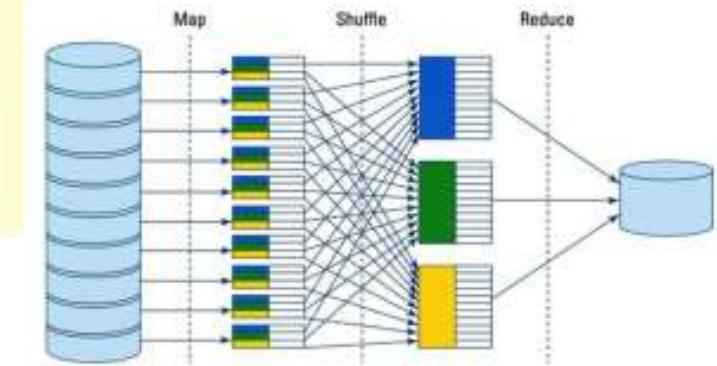


Arquitetura Hadoop



Arquitetura Hadoop

MapReduce foi projetado para usar computação paralela distribuída em Big Data e transformar os dados em pedaços menores.



Arquitetura Hadoop

MapReduce

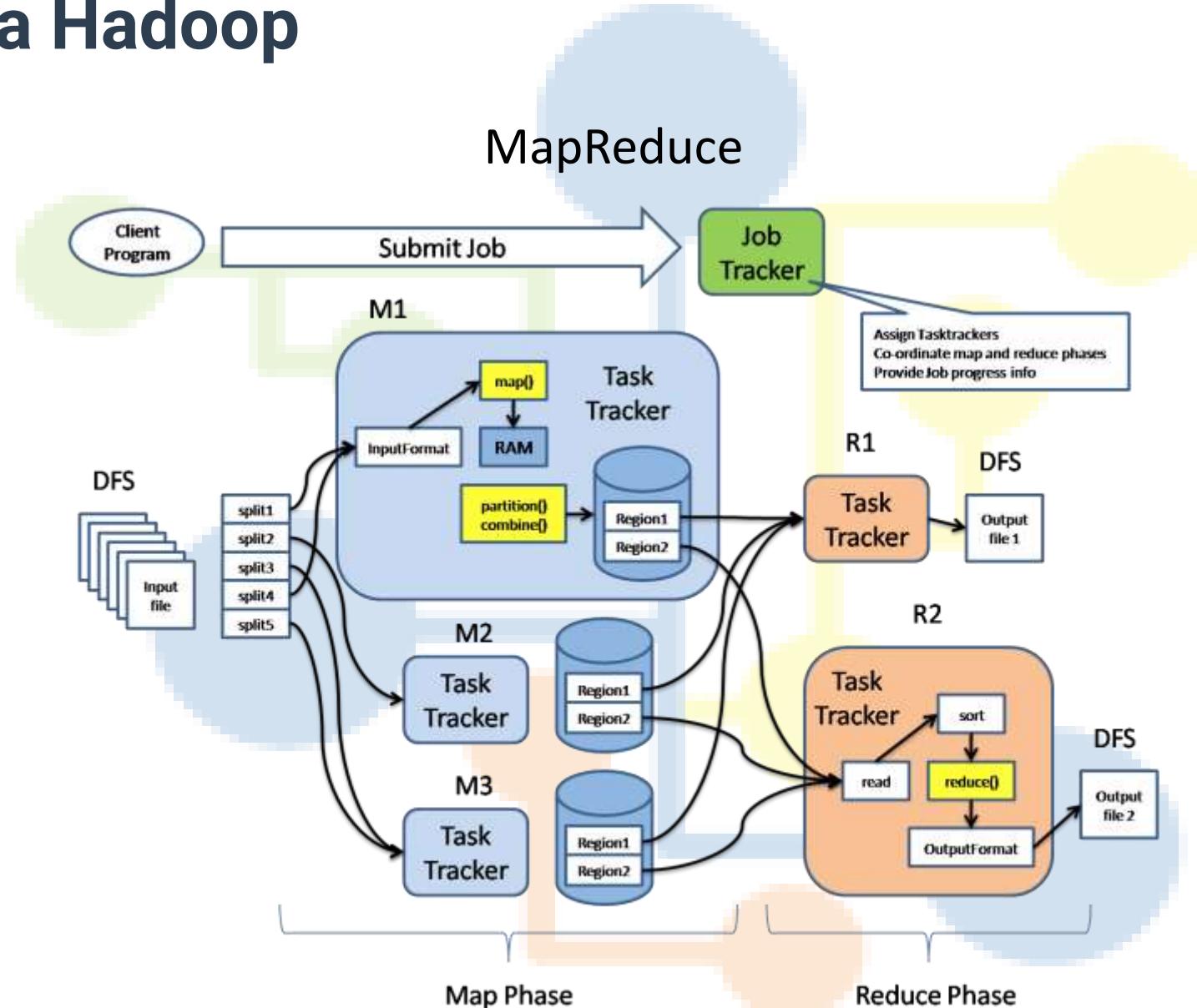
MapReduce funciona através de 2 operações:
Mapeamento e Redução.

No processo de **mapeamento** (Map), os dados são separados em pares (key-value pairs), transformados e filtrados.

Então os dados são distribuídos para os nodes e processados.

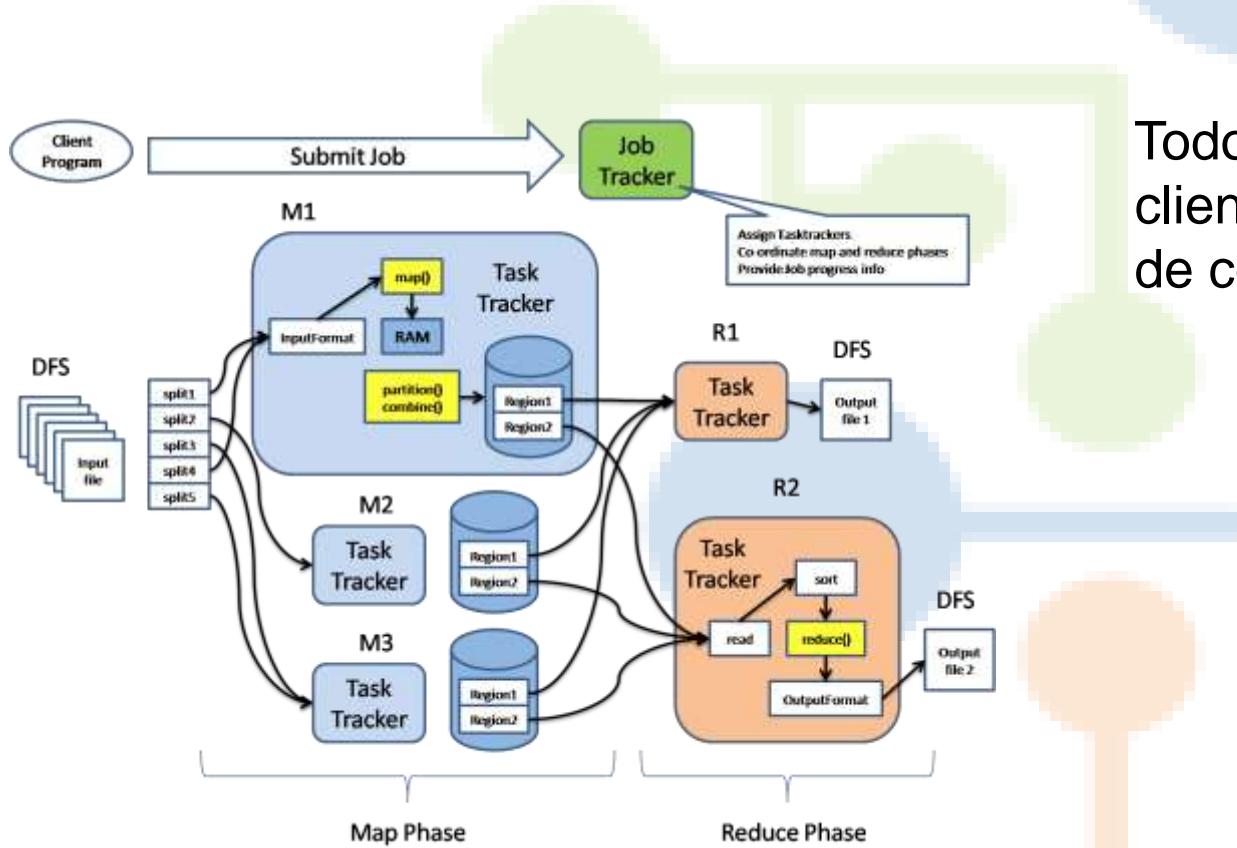
No processo de **redução** (Reduce), os dados são agregados em conjuntos de dados (datasets) menores. Os dados resultantes do processo de redução são transformados em um formato padrão de chave-valor (key-value), onde a chave (key) funciona como o identificador do registro e o valor (value) é o dado (conteúdo) que é identificado pela chave.

Arquitetura Hadoop



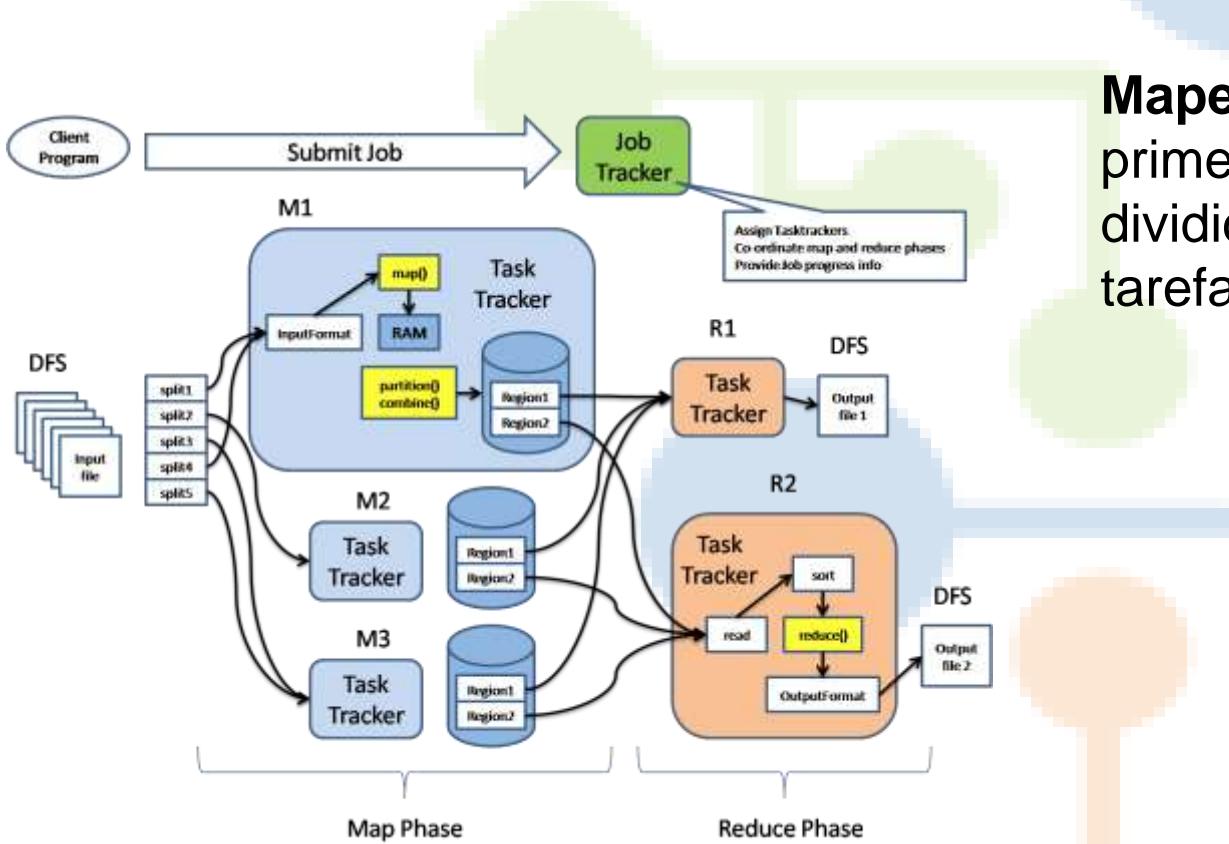
Arquitetura Hadoop

Processo de MapReduce



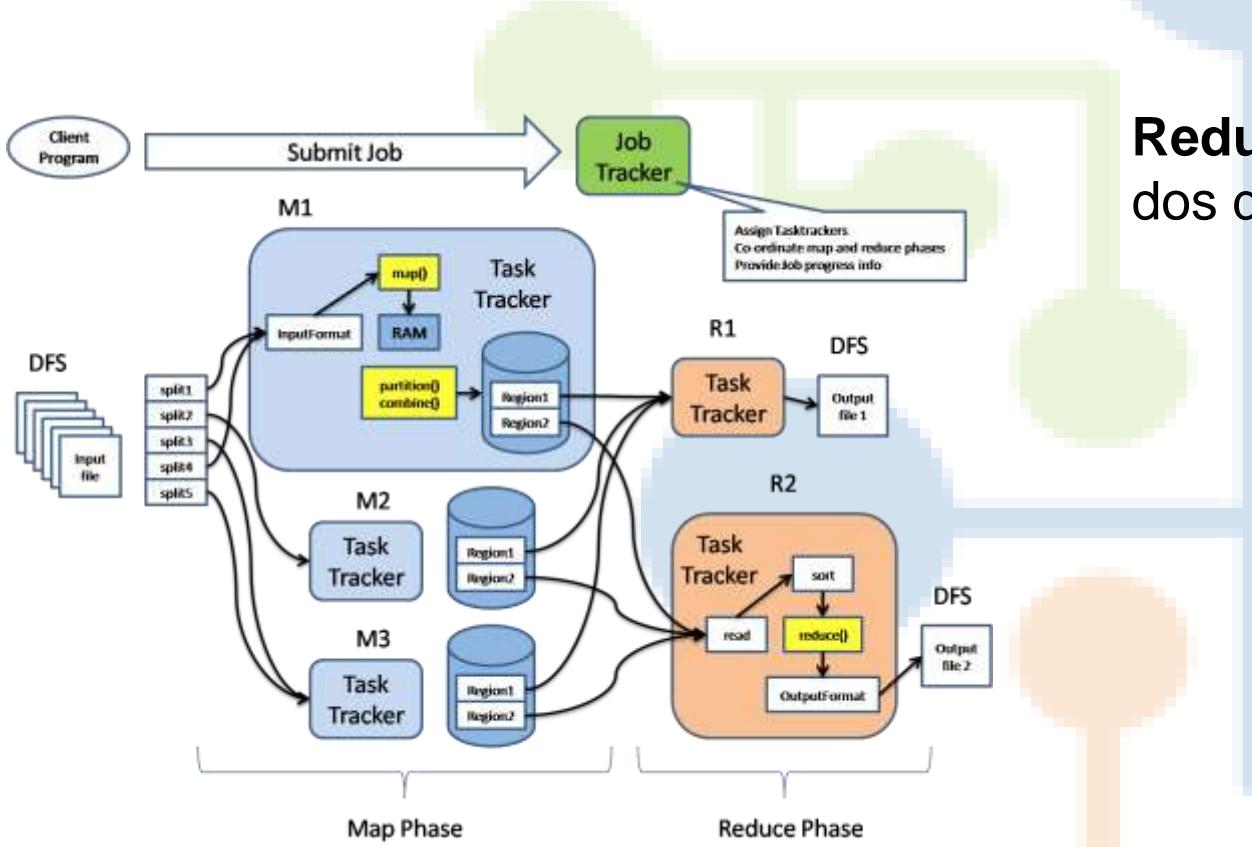
Todo o processo se inicia com a requisição feita pelo cliente e o job submetido. O Job Tracker se encarrega de coordenar como o job será distribuído.

Arquitetura Hadoop



Mapeamento dos dados - os dados de entrada são primeiramente distribuídos em pares key-value e divididos em fragmentos, que são então atribuídos a tarefas de mapeamento.

Arquitetura Hadoop



Redução dos dados - cada operação de redução dos dados tem um fragmento atribuído.

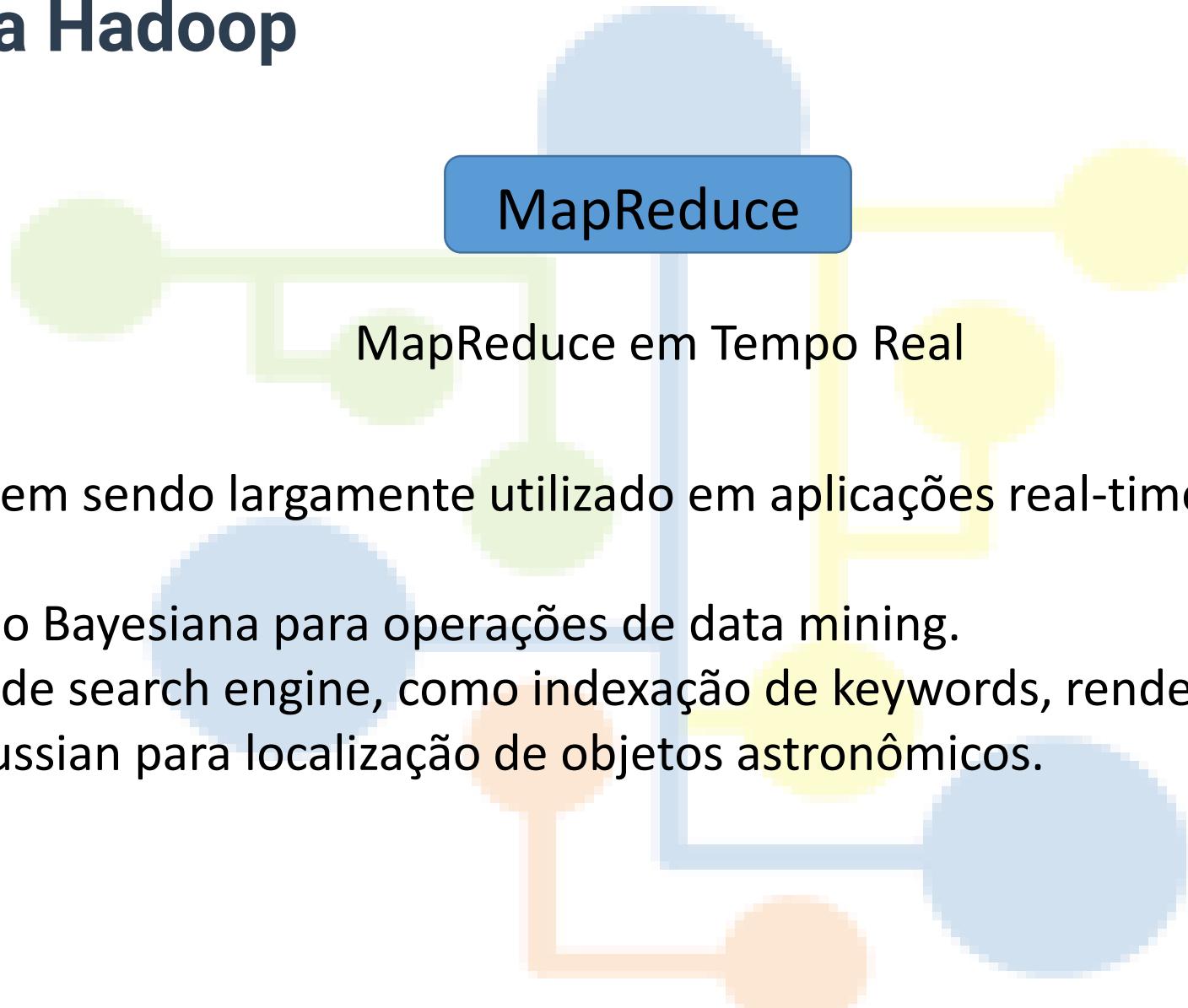
Arquitetura Hadoop

Exemplos de Aplicações do MapReduce

MapReduce vem sendo largamente utilizado em aplicações de Big Data, tais como:

- Classificação Bayesiana para operações de Data Mining.
- Operações de search engine, como indexação de keywords, rendering e page rank.
- Análise Gaussian para localização de objetos astronômicos.
- Web Semântica e Web 3.0.
- Sistemas de Recomendação.

Arquitetura Hadoop



MapReduce vem sendo largamente utilizado em aplicações real-time. Alguns exemplos :

- Classificação Bayesiana para operações de data mining.
- Operações de search engine, como indexação de keywords, rendering e page rank.
- Análise Gaussian para localização de objetos astronômicos.

Arquitetura Hadoop

Cache Distribuído

Distributed Cache ou Cache Distribuído, é uma funcionalidade do Hadoop que permite cache dos arquivos usados pelas aplicações.

Isso permite ganhos consideráveis de performance quando tarefas de map e reduce precisam acessar dados em comum. Permite ainda, que um node do cluster acesse os arquivos no filesystem local, ao invés de solicitar o arquivo em outro node.

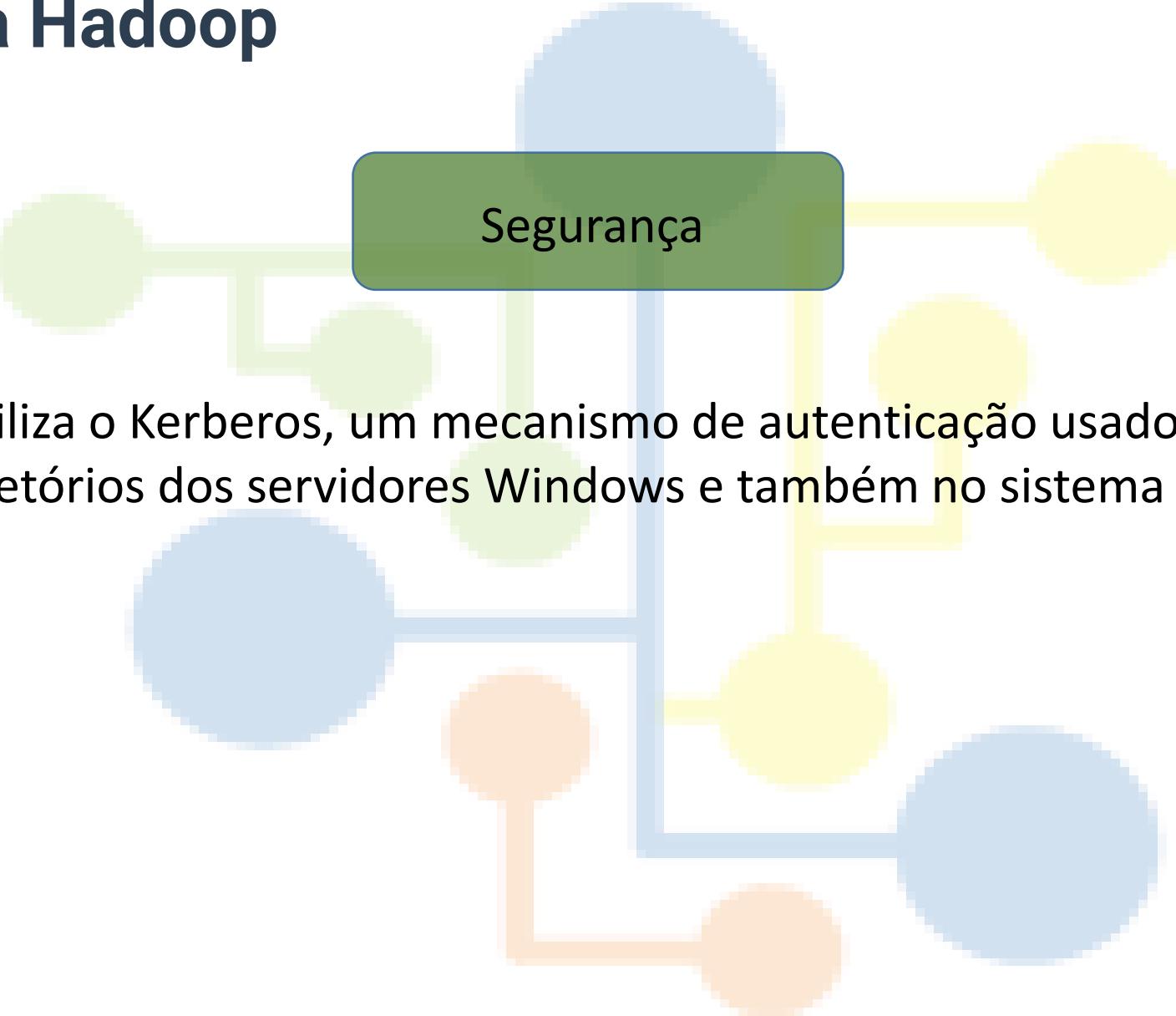
É possível fazer o cache de arquivos zip e tar.gz.

Arquitetura Hadoop

Cache Distribuído

Uma vez que você armazena um arquivo em cache para o seu trabalho, a estrutura Hadoop irá torná-lo disponível em cada node (em sistema de arquivos, não em memória) onde as tarefas de mapeamento / redução estão em execução.

Arquitetura Hadoop



O Hadoop utiliza o Kerberos, um mecanismo de autenticação usado por exemplo no sistema de diretórios dos servidores Windows e também no sistema operacional Linux

Arquitetura Hadoop

Segurança

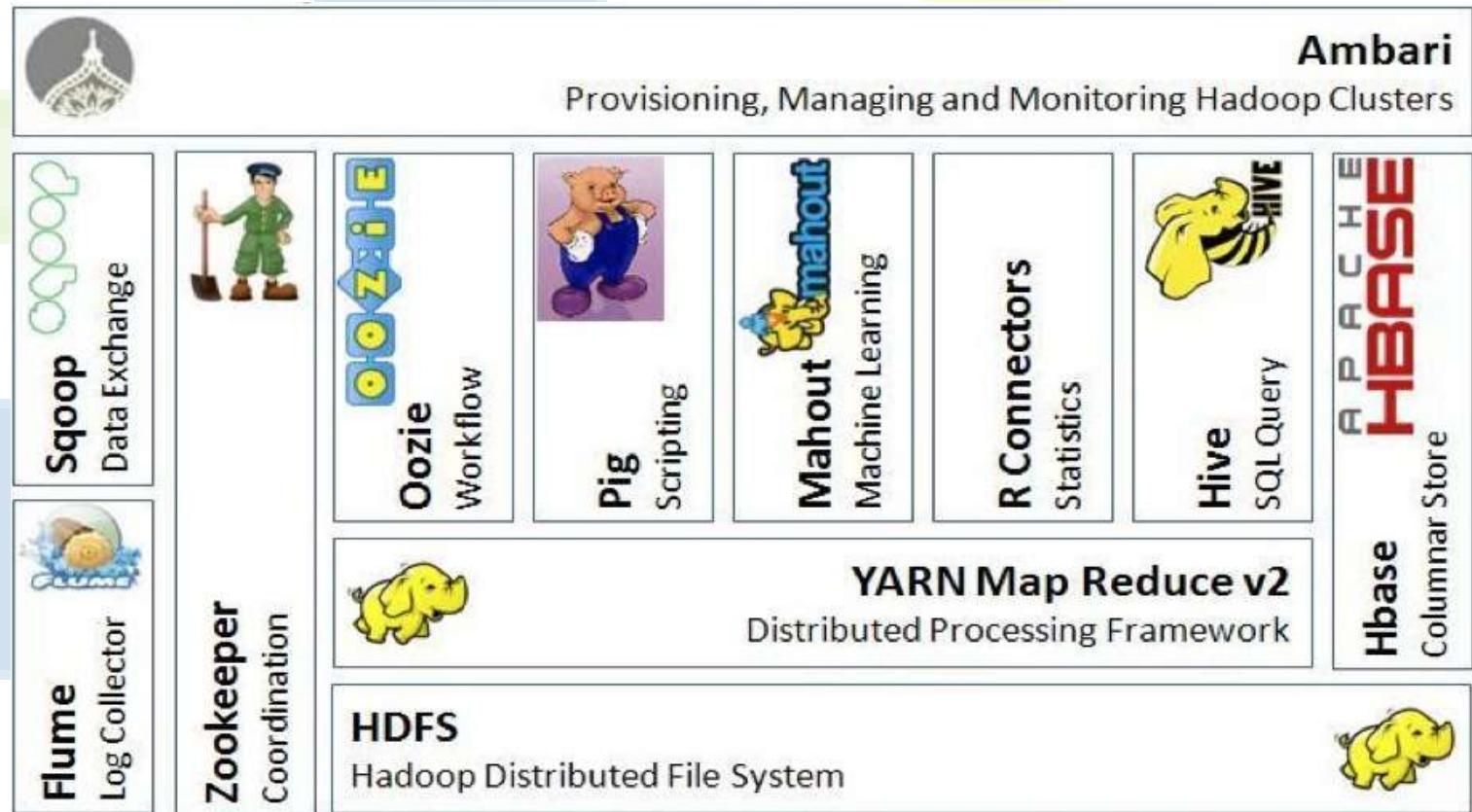
Por padrão Hadoop é executado no modo não-seguro em que não é necessária a autenticação real. Após ser configurado, o Hadoop é executado em modo de segurança e cada usuário e serviço precisa ser autenticado pelo Kerberos, a fim de utilizar os serviços do Hadoop.

Arquitetura Hadoop

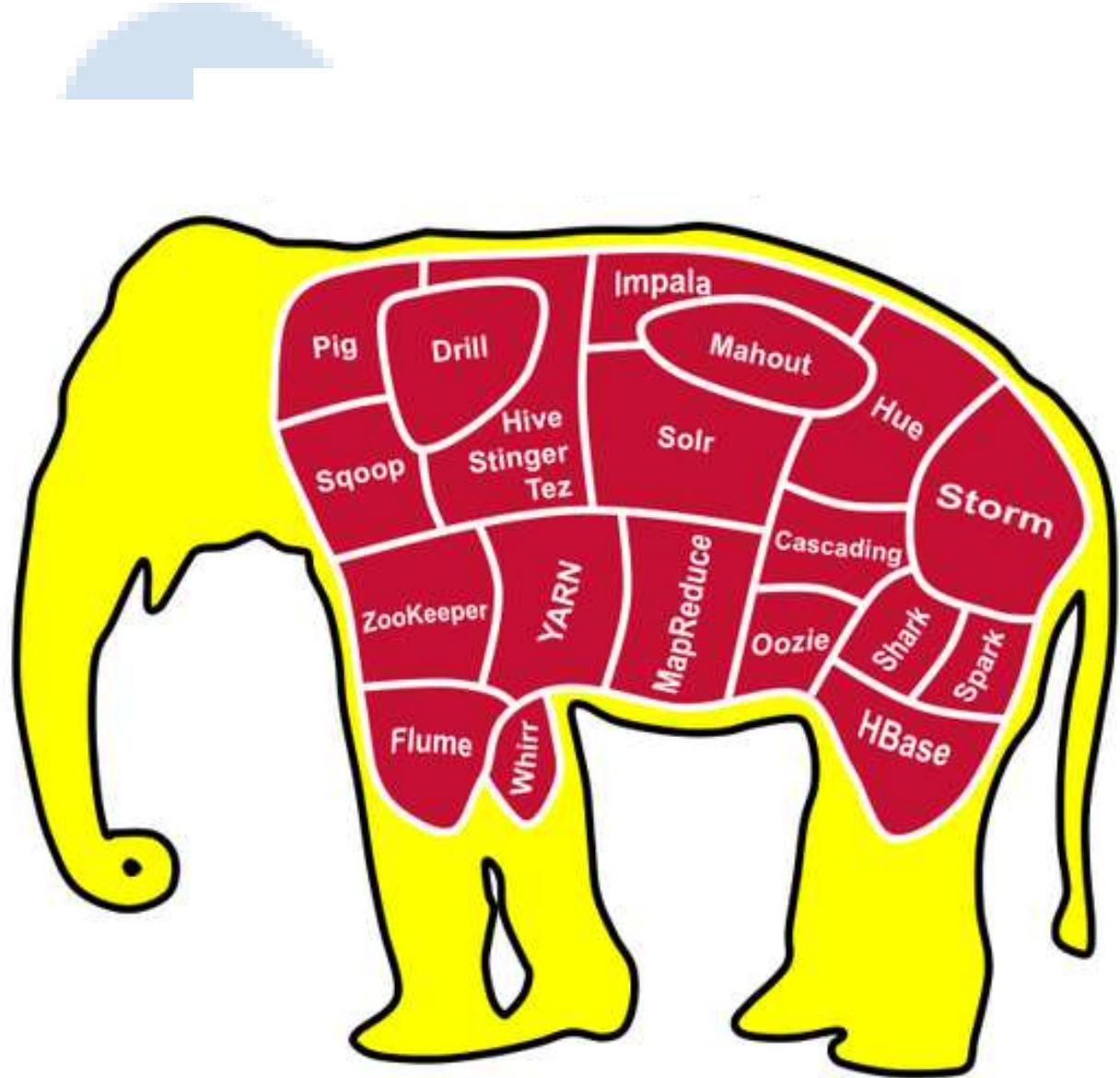
Segurança

Depois que o Kerberos estiver configurado, a autenticação Kerberos é usada para validar as credenciais do lado do cliente. Isso significa que o cliente deve solicitar uma permissão de serviço válido para o ambiente Hadoop.

Ecossistema Hadoop



Ecossistema Hadoop



Ecossistema Hadoop

Pense no ecossistema como as apps do sistema operacional iOS ou Android

Os aplicativos servem para aprimorar a capacidade do SO

Mesmo raciocínio pode ser aplicado para os componentes do ecossistema Hadoop

Ecossistema Hadoop



Ecossistema Hadoop

Apache Zookeeper



<http://zookeeper.apache.org>

Zookeeper é uma solução open-source de alta performance, para coordenação de serviços em aplicações distribuídas.

Ele é uma espécie de guardião do Zoo!

Ecossistema Hadoop

Apache Zookeeper



<http://zookeeper.apache.org>

ZooKeeper é um serviço de coordenação distribuída para gerenciar grandes conjuntos de hosts (Clusters).



Ecossistema Hadoop

Apache Zookeeper



<http://zookeeper.apache.org>

Coordenação e gestão de um serviço em um ambiente distribuído é um processo complicado.

ZooKeeper resolve este problema com a sua arquitetura simples.

Ecossistema Hadoop

Apache Zookeeper



<http://zookeeper.apache.org>

ZooKeeper permite que os desenvolvedores se concentrem na lógica do aplicativo principal sem se preocupar com a natureza distribuída do aplicativo.

Ecossistema Hadoop

Apache Zookeeper



<http://zookeeper.apache.org>

O framework ZooKeeper foi originalmente construído no "Yahoo!" para acessar seus aplicativos de uma forma fácil e robusta.

Mais tarde, Apache ZooKeeper se tornou um padrão para a organização de serviços do Hadoop, HBase e outras estruturas distribuídas.

Por exemplo, o HBase usa ZooKeeper para acompanhar o estado de dados distribuídos através do Cluster.

Ecossistema Hadoop

Apache Zookeeper



<http://zookeeper.apache.org>

ZooKeeper proporciona um ponto comum de acesso a uma ampla variedade de objetos utilizados em ambientes de Cluster.

Ecossistema Hadoop

Apache Oozie



<http://oozie.apache.org>

Apache Oozie é um sistema de agendamento de workflow usado para gerenciar principalmente os Jobs de MapReduce.

Ecossistema Hadoop

Apache Oozie



<http://oozie.apache.org>

Oozie é integrado com o restante dos componentes do ecossistema Hadoop para apoiar vários tipos de trabalhos do Hadoop (como Java Map-Reduce, streaming Map-Reduce, Pig, Hive e Sqoop), bem como jobs específicos do sistema (como programas Java e scripts shell).

Ecossistema Hadoop

Apache Oozie



<http://oozie.apache.org>

Oozie é um sistema de processamento de fluxo de trabalho que permite aos usuários definir uma série de jobs escritos em diferentes linguagens - como Map Reduce, Pig e Hive – e então inteligentemente ligá-los um ao outro.

Ecossistema Hadoop

Apache Oozie



<http://oozie.apache.org>

Oozie permite aos usuários especificar, por exemplo, que uma determinada consulta só pode ser iniciada, após os jobs anteriores que accessem os mesmos dados, sejam concluídos.

Ecossistema Hadoop

Apache Oozie



<http://oozie.apache.org>

Oozie é um sistema versátil que pode ser usado para configurar e automatizar até mesmo o mais complicado workflow de processamento de dados.

Lembre-se que estamos falando em processamento de Big Data, em Clusters que podem chegar a milhares de nodes.

Ecossistema Hadoop

Apache Hive



<http://hive.apache.org>

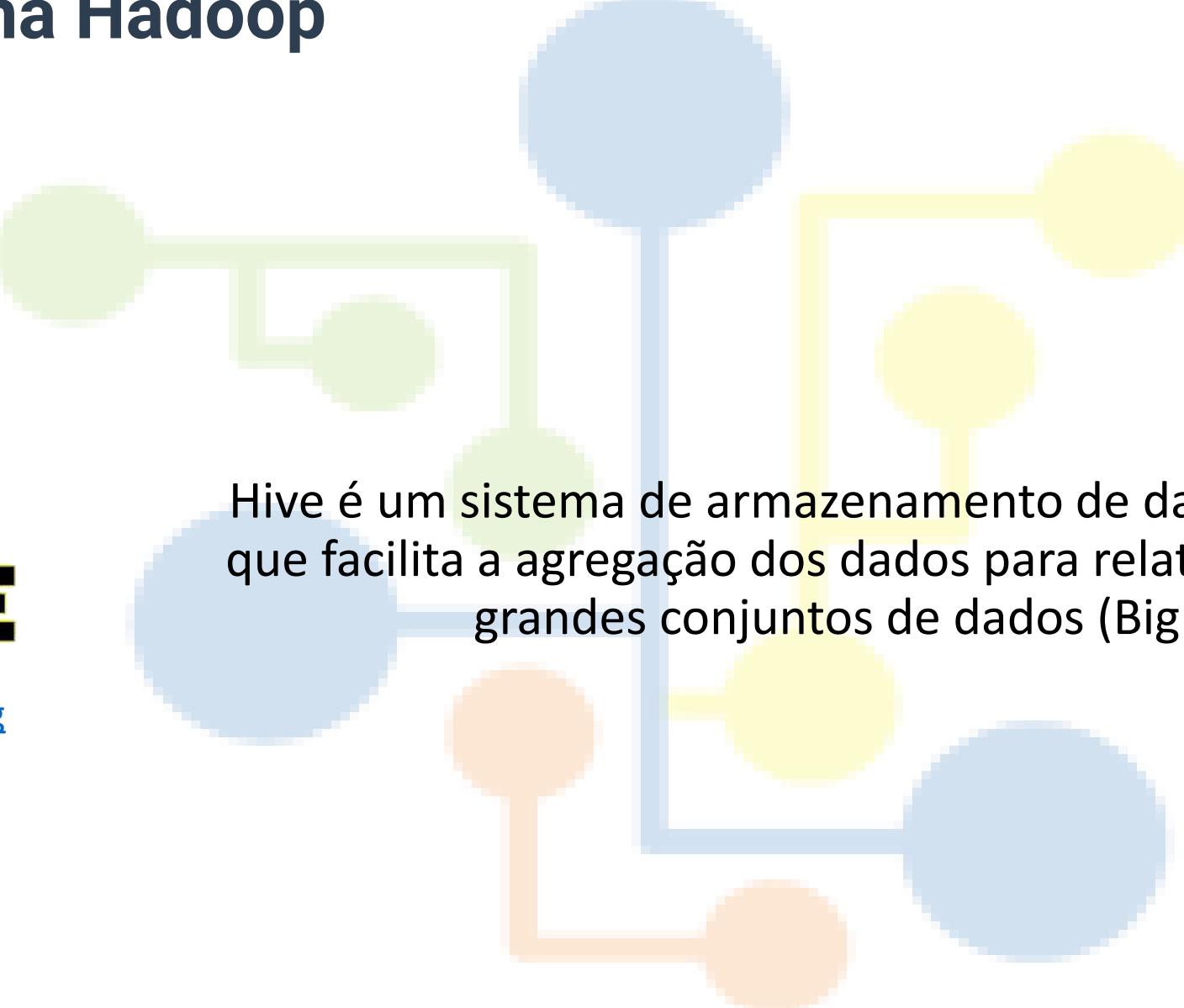
Apache Hive é um Data Warehouse que funciona com Hadoop e MapReduce.

Ecossistema Hadoop

Apache Hive



<http://hive.apache.org>



Hive é um sistema de armazenamento de dados para Hadoop que facilita a agregação dos dados para relatórios e análise de grandes conjuntos de dados (Big Data).

Ecossistema Hadoop

Apache Hive



<http://hive.apache.org>

Hive permite consultas sobre os dados usando uma linguagem SQL-like, chamada HiveQL (HQL).

Ecossistema Hadoop

Apache Hive



<http://hive.apache.org>

Provê capacidade de tolerância a falha para armazenamento de dados e depende do MapReduce para execução.

Ecossistema Hadoop

Apache Hive



<http://hive.apache.org>

Ele permite conexões JDBC / ODBC, por isso é facilmente integrado com outras ferramentas de inteligência de negócios como Tableau, Microstrategy, Microsoft Power BI entre outras.

Ecossistema Hadoop

Apache Hive



<http://hive.apache.org>

Hive é orientado a batch e possui alta latência para execução de queries.

Assim como o Pig, gera jobs MapReduce que executam no cluster Hadoop.

Ecossistema Hadoop

Apache Hive



<http://hive.apache.org>

Hive é um sistema para gestão e query de dados não estruturados, em formato estruturado.

Hive utiliza:

MapReduce
(para execução)

HDFS
(para armazenamento
e pesquisa de dados)

Ecossistema Hadoop

Apache Hive



<http://hive.apache.org>

Hive Query Language - HQL

Hive Query Language (HQL) é a linguagem de queries para o engine Hive

HQL suporta os conceitos básicos da linguagem SQL

- Cláusula From
- ANSI Join (somente equi-join)
- Insert
- Group-by
- Sampling

Ecossistema Hadoop



<http://hive.apache.org>

Hive Query Language - HQL

Exemplo:

```
hive> select * from tb_folha_pagamento;
```

```
hive> show tables;
```

```
hive> describe tb_folha_pagamento;
```

Ecossistema Hadoop

Apache Sqoop



<http://sqoop.apache.org>

Sqoop é um projeto do ecossistema do Apache Hadoop, cuja responsabilidade é importar e exportar dados de bancos de dados relacionais.

Ecossistema Hadoop

Apache Sqoop



<http://sqoop.apache.org>

Sqoop é um projeto do ecossistema do Apache Hadoop, cuja responsabilidade é importar e exportar dados de bancos de dados relacionais.

Sqoop significa SQL-to-Hadoop.

Ecossistema Hadoop

Apache Soop



<http://sqoop.apache.org>

Basicamente, o Soop permite mover os dados de bancos tradicionais como Microsoft SQL Server ou Oracle, para o Hadoop.

Ecossistema Hadoop

Apache Sqoop



<http://sqoop.apache.org>

É possível importar tabelas individuais ou bancos de dados inteiros para o HDFS e o desenvolvedor pode determinar que colunas ou linhas serão importadas.

Ecossistema Hadoop

Apache Sqoop



<http://sqoop.apache.org>

Ferramenta desenvolvida para transferir dados do Hadoop para RDBMS e vice-versa.

Ecossistema Hadoop

Apache Sqoop



<http://sqoop.apache.org>

Transforma os dados no Hadoop, sem necessidade de desenvolvimento adicional.



Ecossistema Hadoop

Apache Sqoop



<http://sqoop.apache.org>

Ele também gera classes Java através das quais você pode facilmente interagir com os dados importados.

Ecossistema Hadoop

Apache Sqoop



<http://sqoop.apache.org>

Utiliza conexão JDBC para conectar com os bancos de dados relacionais.

Ecossistema Hadoop

Apache Sqoop



<http://sqoop.apache.org>

Pode criar diretamente tabelas no Hive e suporta importação incremental.

Ecossistema Hadoop

Apache Soop



<http://sqoop.apache.org>

Exemplo: Listando tabelas de um banco MySQL com Soop:

```
sqoop list-tables --username dsacademy --password dsacademybr \
--connect jdbc:mysql://dbname
```

Ecossistema Hadoop

Apache Pig



<http://pig.apache.org>

É uma ferramenta que é utilizada para analisar grandes conjuntos de dados que representam fluxos de dados.

Ecossistema Hadoop

Apache Pig



<http://pig.apache.org>

Podemos realizar todas as operações de manipulação de dados no Hadoop usando Apache Pig.

Ecossistema Hadoop

Apache Pig



<http://pig.apache.org>

Para escrever programas de análise de dados, Pig oferece uma linguagem de alto nível conhecida como Pig Latin.

Ecossistema Hadoop

Apache Pig



<http://pig.apache.org>

Para escrever programas de análise de dados, Pig oferece uma linguagem de alto nível conhecida como Pig Latin.

Esta linguagem fornece vários operadores que os programadores podem usar para criar suas próprias funções para leitura, escrita e processamento de dados.

Ecossistema Hadoop

Apache Pig



<http://pig.apache.org>

Para analisar dados usando Apache Pig, os programadores precisam escrever scripts usando linguagem Pig Latin.

Ecossistema Hadoop

Apache Pig



<http://pig.apache.org>

Para analisar dados usando Apache Pig, os programadores precisam escrever scripts usando linguagem Pig Latin.

Todos esses scripts são convertidos internamente para tarefas de mapeamento e redução.

Ecossistema Hadoop

Apache Pig



<http://pig.apache.org>

Para analisar dados usando Apache Pig, os programadores precisam escrever scripts usando linguagem Pig Latin.

Todos esses scripts são convertidos internamente para tarefas de mapeamento e redução.

Apache Pig tem um componente conhecido como Pig engine que aceita os scripts Pig Latin como entrada e converte esses scripts em jobs MapReduce.

Ecossistema Hadoop

Apache Pig



<http://pig.apache.org>

Componentes do Pig

→ Pig Latin Script Language

- Linguagem procedural de fluxo de dados
- Contém sintaxe e comandos que podem ser aplicados para implementar lógica de negócios

→ Runtime engine

- Compilador que produz sequências de programas MapReduce
- Utiliza HDFS para armazenar e buscar dados
- Usado para interagir com sistemas Hadoop
- Valida e compila scripts em sequências de Jobs MapReduce

Ecossistema Hadoop

Apache Pig



<http://pig.apache.org>

Pig X SQL

| Pig | SQL |
|---|--|
| Linguagem de script usada para interagir com o HDFS | Linguagem de query usada para interagir com bancos de dados |
| Passo a passo | Bloco único |
| Avaliação não imediata | Avaliação imediata |
| Permite resultados intermediários | Requer que um join seja executado 2 vezes ou materializado como um resultado intermediário |

Ecossistema Hadoop

Apache HBase



<http://hbase.apache.org>

HBase é um banco de dados orientado a coluna construído sobre o sistema de arquivos do Hadoop.

Ecossistema Hadoop

Apache HBase



<http://hbase.apache.org>

HBase é um banco de dados orientado a coluna construído sobre o sistema de arquivos do Hadoop.

HBase é o banco de dados oficial do Hadoop.

Ecossistema Hadoop

Apache HBase



<http://hbase.apache.org>

HBase tem um modelo de dados que é semelhante ao Big Table do Google projetado para fornecer acesso aleatório rápido a grandes quantidades de dados.

Ecossistema Hadoop

Apache HBase



<http://hbase.apache.org>

Ele aproveita a tolerância a falhas fornecida pelo sistema de arquivos do Hadoop (HDFS).

Ecossistema Hadoop

Apache HBase



<http://hbase.apache.org>

Ele aproveita a tolerância a falhas fornecida pelo sistema de arquivos do Hadoop (HDFS).

É uma parte do ecossistema Hadoop que fornece em tempo real acesso aleatório de leitura / gravação aos dados do HDFS.

Ecossistema Hadoop

Apache HBase



<http://hbase.apache.org>

Podemos armazenar os dados diretamente no HDFS ou através do HBase.

Ecossistema Hadoop

Apache HBase



<http://hbase.apache.org>

O objetivo do HBase é armazenar tabelas realmente grandes, com bilhões de registros.

Ecossistema Hadoop

Apache HBase



<http://hbase.apache.org>

Arquitetura HBase

HBase possui 2 tipos de Nodes: Master e RegionServer

| Master | RegionServer |
|--|---|
| Somente um node Master pode ser executado. A alta disponibilidade é mantida pelo ZooKeeper | Um ou mais podem existir |
| Responsável pela gestão de operações de cluster, como assignment, load balancing e splitting | Responsável por armazenar as tabelas, realizar leituras e buffers de escrita |
| Não faz parte de operações de read/write | O cliente comunica com o RegionServer para processar operações de leitura/escrita |

Ecossistema Hadoop

Apache HBase



<http://hbase.apache.org>

HBase x RDBMS

| HBase | RDBMS |
|--|--|
| Particionamento automático | Particionamento automático ou manual, realizado pelo administrador |
| Pode ser escalado de forma linear e automática com novos nodes | Pode ser escalado verticalmente com a adição de mais hardware |
| Utiliza hardware commodity | Requer hardware mais robusto e portanto, mais caro |
| Possui tolerância a falha | Tolerância a falha pode estar presente ou não |

Ecossistema Hadoop

Apache Flume



<http://flume.apache.org>

Flume é um serviço que basicamente permite enviar dados diretamente para o HDFS.



Ecossistema Hadoop

Apache Flume



<http://flume.apache.org>

Foi desenvolvido pela Cloudera e permite mover grandes quantidades de dados.

Ecossistema Hadoop

Apache Flume



<http://flume.apache.org>

Basicamente, o Apache Flume é um serviço que funciona em ambiente distribuído para coletar, agrregar e mover grandes quantidades de dados de forma eficiente.

Ecossistema Hadoop

Apache Flume



<http://flume.apache.org>

Ele possui uma arquitetura simples e flexível baseada em streaming (fluxo constante) de dados.

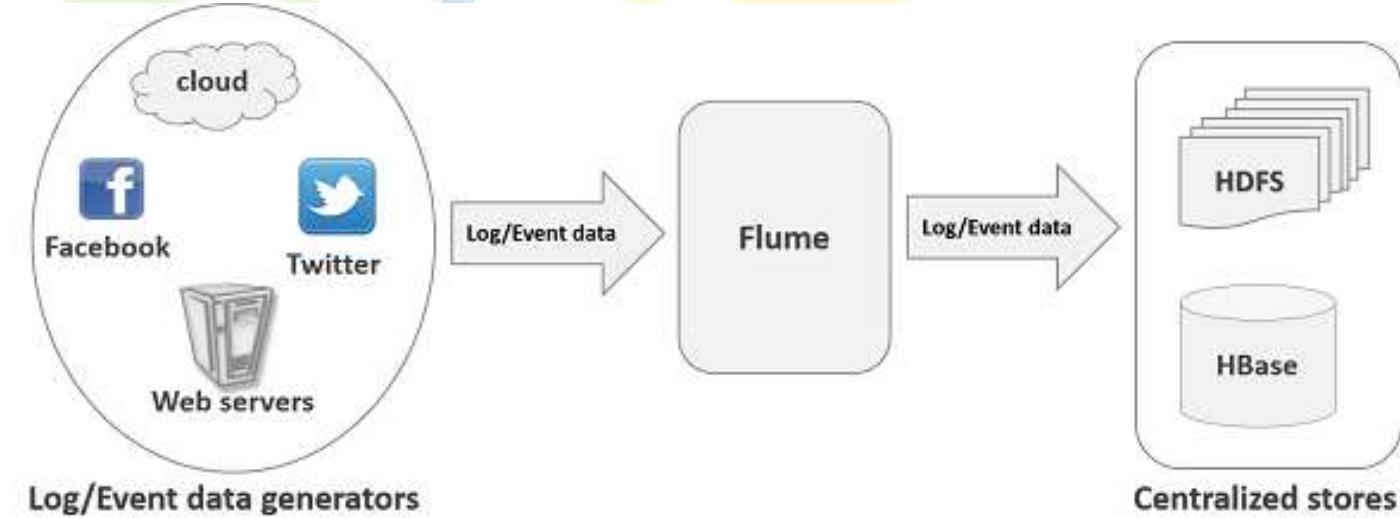
Ecossistema Hadoop

Apache Flume



<http://flume.apache.org>

Ele possui uma arquitetura simples e flexível baseada em streaming (fluxo constante) de dados.



Ecossistema Hadoop

Apache Flume



<http://flume.apache.org>

O modelo de dados do Flume, permite que ele seja usado em aplicações analíticas online.

Ecossistema Hadoop

Apache Flume



<http://flume.apache.org>

O Flume também pode ser usado em Infraestrutura de TI.



Ecossistema Hadoop

Apache Flume



<http://flume.apache.org>

O Flume também pode ser usado em Infraestrutura de TI.

Agentes são instalados em servidores web, servidores de aplicação ou aplicativos mobile, para coletar e integrar os dados com Hadoop, para análise online.

Ecossistema Hadoop

Apache Mahout



<http://mahout.apache.org>

Apache Mahout é uma biblioteca open-source de algoritmos de aprendizado de máquina, escalável e com foco em clustering, classificação e sistemas de recomendação.

Ecossistema Hadoop



<http://mahout.apache.org>

O Mahout é dedicado ao Machine Learning.

Ecossistema Hadoop

Apache Mahout



<http://mahout.apache.org>

O Mahout permite a utilização dos principais algoritmos de clustering, testes de regressão e modelagem estatística e os implementa usando um modelo MapReduce.

Ecossistema Hadoop



<http://mahout.apache.org>

Apache Mahout

E quando utilizar o Mahout?

Ecossistema Hadoop

Apache Mahout



<http://mahout.apache.org>

- Você precisa utilizar algoritmos de Machine Learning com alta performance?
- Sua solução precisa ser open-source e gratuita?
- Você possui um grande conjunto de dados (Big Data) e pretende utilizar ferramentas de análise como R, Python e Octave?
- Seu processamento de dados será feito usando um modelo batch (você não precisa utilizar dados gerados em tempo real)?

Ecossistema Hadoop

Apache Mahout



<http://mahout.apache.org>

- Você precisa utilizar algoritmos de Machine Learning com alta performance?
- Sua solução precisa ser open-source e gratuita?
- Você possui um grande conjunto de dados (Big Data) e pretende utilizar ferramentas de análise como R, Python e Octave?
- Seu processamento de dados será feito usando um modelo batch (você não precisa utilizar dados gerados em tempo real)?
- Você precisa de uma biblioteca madura e disponível no mercado há alguns anos que já tenha sido testada e validada?

Ecossistema Hadoop

Apache Mahout



<http://mahout.apache.org>

Se suas respostas forem sim, o Mahout pode atender suas necessidades.

Ecossistema Hadoop

Apache Kafka



<http://kafka.apache.org>

O Apache Kafka foi desenvolvido pelo LinkedIn e posteriormente liberado como um projeto open-source, em 2011.

Ecossistema Hadoop

Apache Kafka



<http://kafka.apache.org>

O Apache Kafka é um sistema para gerenciamento de fluxos de dados em tempo real, gerados a partir de web sites, aplicações e sensores.

Ecossistema Hadoop

Apache Kafka



<http://kafka.apache.org>

Essencialmente, o Kafka age como uma espécie de “sistema nervoso central”, que coleta dados de alto volume como por exemplo a atividade de usuários (clicks em um web site), logs, cotações de ações etc... e torna estes dados disponíveis como um fluxo em tempo real para o consumo por outras aplicações.

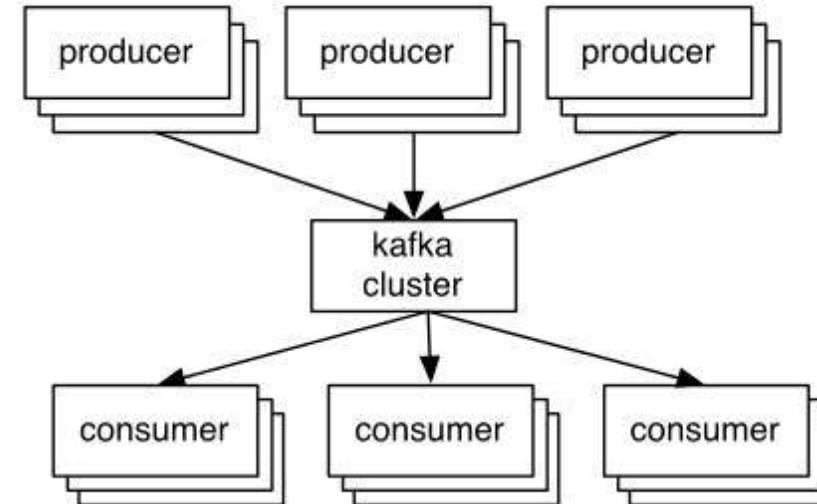
Ecossistema Hadoop

Apache Kafka



<http://kafka.apache.org>

Apache Kafka



Ecossistema Hadoop

Apache Kafka



O Apache Kafka foi desenvolvido com um propósito específico em mente: servir como um repositório central de fluxos de dados.

<http://kafka.apache.org>

Ecossistema Hadoop

Apache Kafka



<http://kafka.apache.org>

O Apache Kafka foi desenvolvido com um propósito específico em mente: servir como um repositório central de fluxos de dados.

Mas por que fazer isso?

Ecossistema Hadoop

Apache Kafka



<http://kafka.apache.org>

O Apache Kafka foi desenvolvido com um propósito específico em mente: servir como um repositório central de fluxos de dados

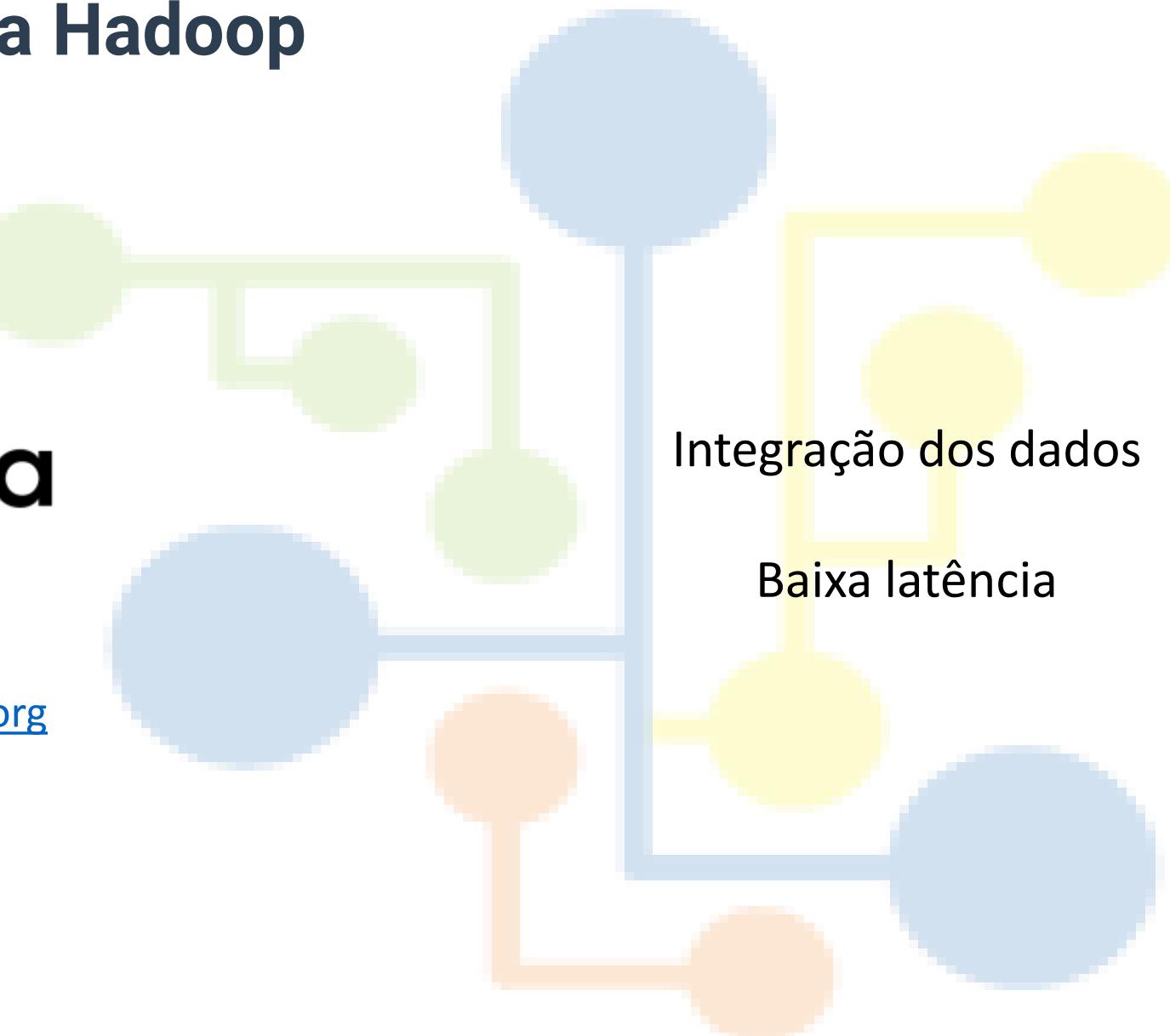
Mas por que fazer isso?

Havia duas motivações.

Ecossistema Hadoop



<http://kafka.apache.org>



Ecossistema Hadoop

Apache Kafka



<http://kafka.apache.org>

O Apache Kafka está ajudando a mudar a forma como os dados são usados dentro das empresas.

Não faz mais sentido falar apenas em dados armazenados em tabelas, com linhas e colunas.

Ecossistema Hadoop

Apache Kafka



<http://kafka.apache.org>

O Apache Kafka está ajudando a mudar a forma como os dados são usados dentro das empresas.

Não faz mais sentido falar apenas em dados armazenados em tabelas, com linhas e colunas.

O volume de dados agora é tão grande, que os dados precisam ser vistos como o que realmente são: um fluxo constante, que precisa ser analisado em tempo real.



Soluções Comerciais com Hadoop

Soluções Comerciais com Hadoop

Por que usar soluções comerciais com Hadoop



Soluções Comerciais com Hadoop

Por que usar soluções comerciais com Hadoop?

Você pode estar se perguntando:

Se o Hadoop é livre, porque eu usaria soluções comerciais do software?

- **Suporte** – as principais soluções comerciais do Hadoop oferecem suporte, guias, assistência e melhores práticas.
- **Confiança** – sempre que um bug é detectado, as soluções comerciais prontamente atualizam o software.
- **Pacote completo** – as soluções oferecem pacotes completos, com tudo que é necessário para uma infraestrutura de Big Data.

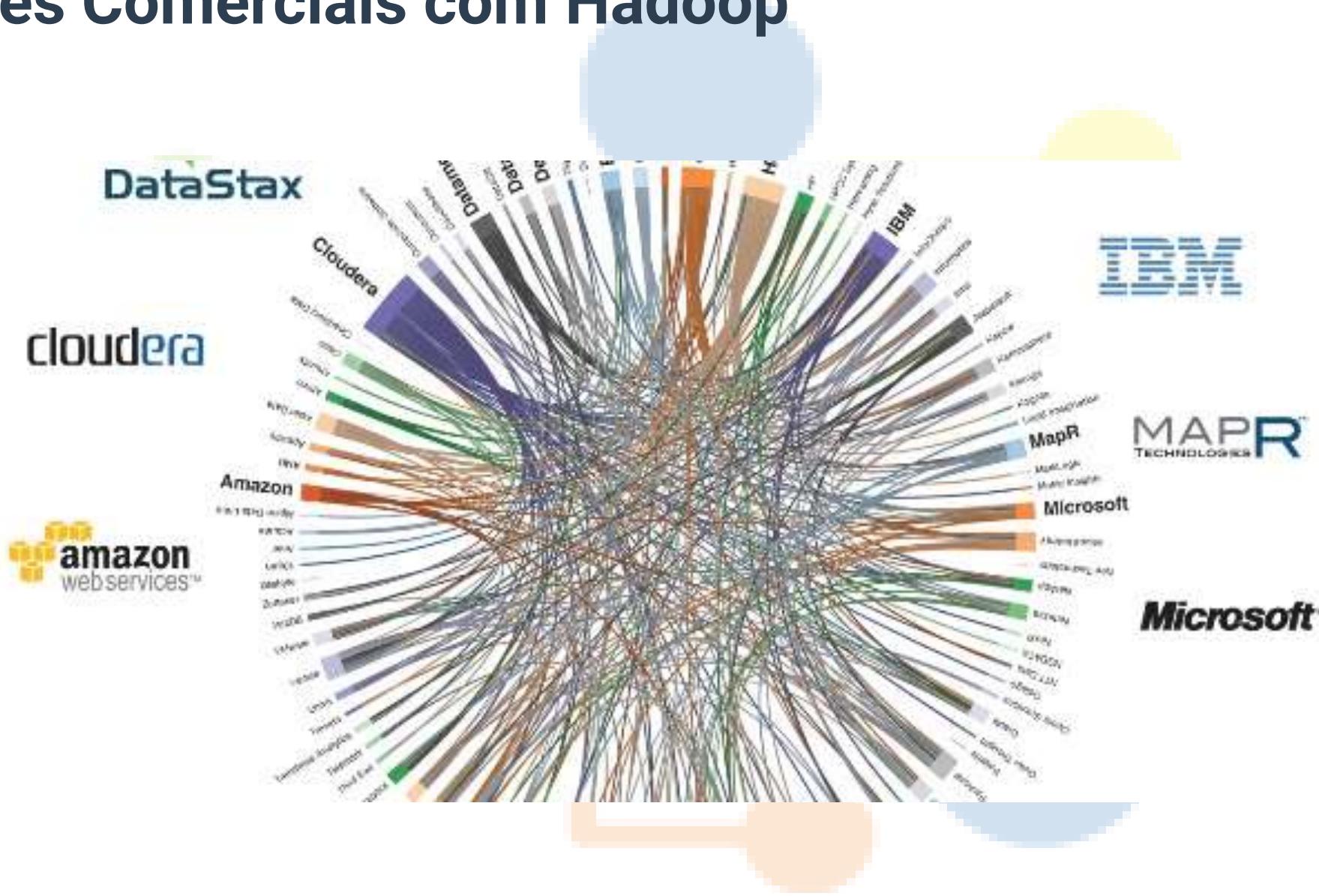


Soluções Comerciais com Hadoop

Principais Soluções Comerciais com Hadoop

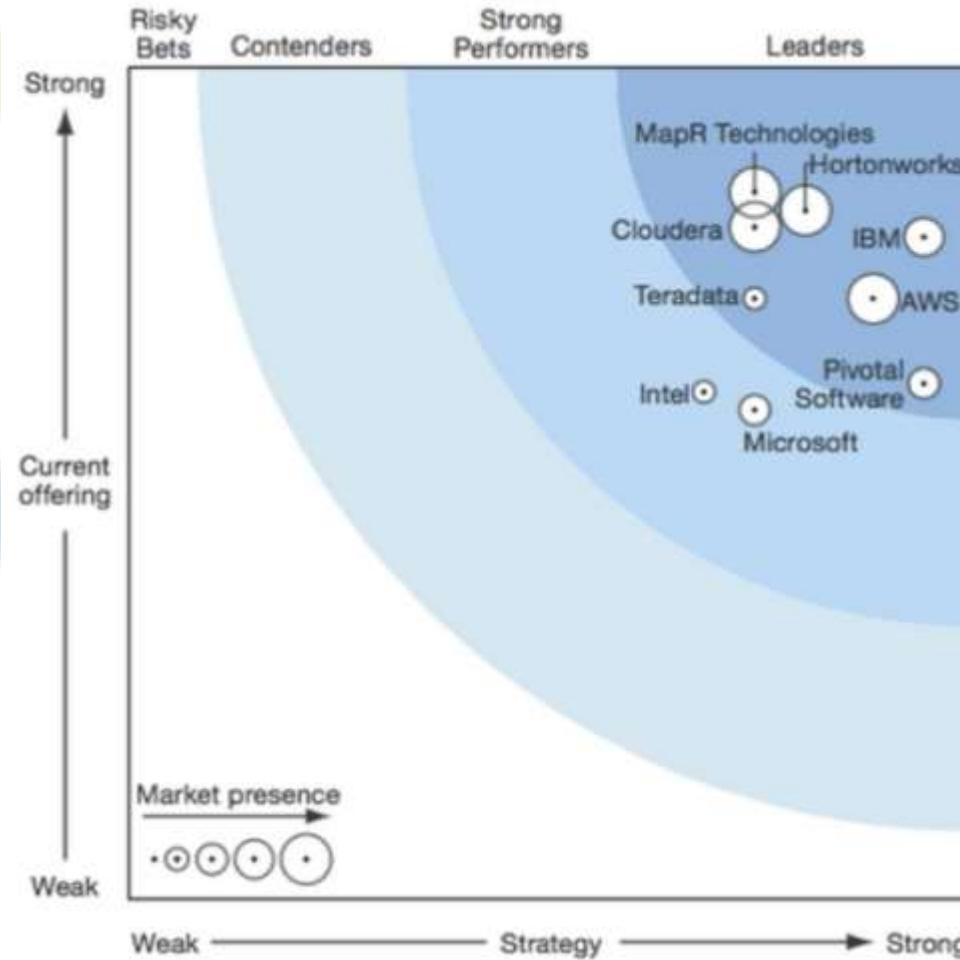


Soluções Comerciais com Hadoop

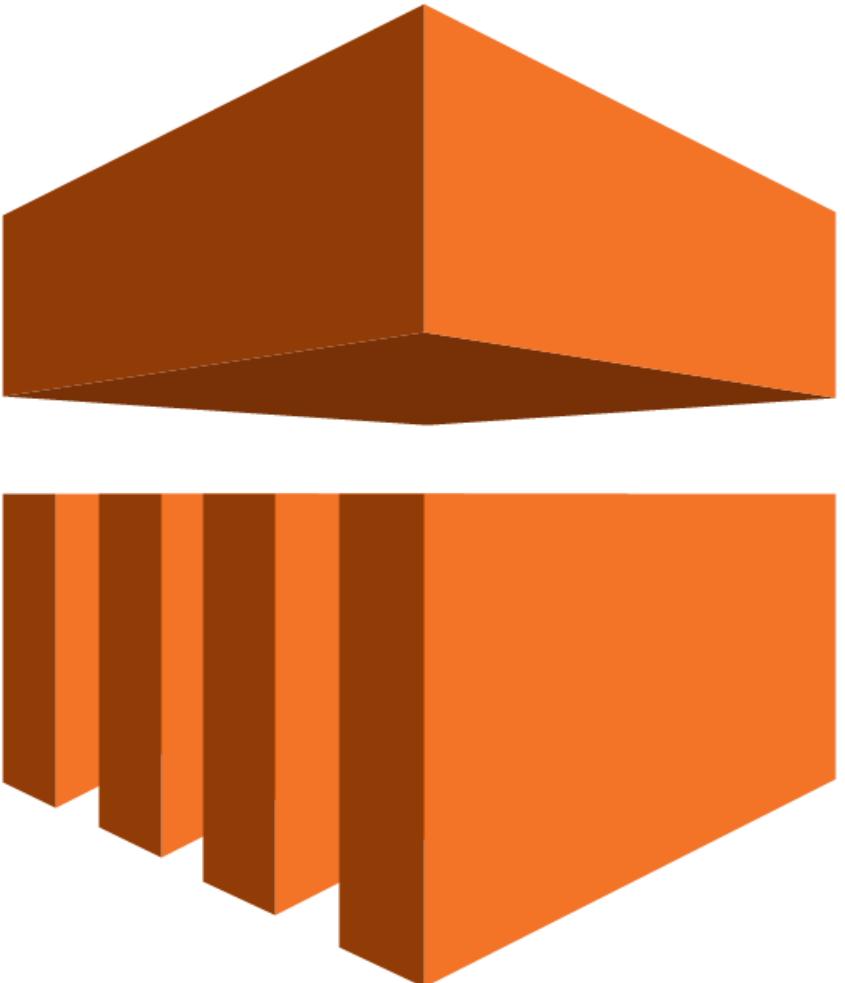


Soluções Comerciais com Hadoop

Principais Soluções Comerciais com Hadoop



Soluções Comerciais com Hadoop



Soluções Comerciais com Hadoop



A distribuição Hadoop da Amazon, foi uma das primeiras distribuições comerciais do Hadoop

Soluções Comerciais com Hadoop



AWS Elastic MapReduce é uma plataforma de análise de dados bem organizada e construída sobre a arquitetura HDFS

Soluções Comerciais com Hadoop



Com foco principal em consultas de mapeamento / redução o AWS EMR explora ferramentas Hadoop, fornecendo uma plataforma de infraestrutura escalável e segura para seus usuários

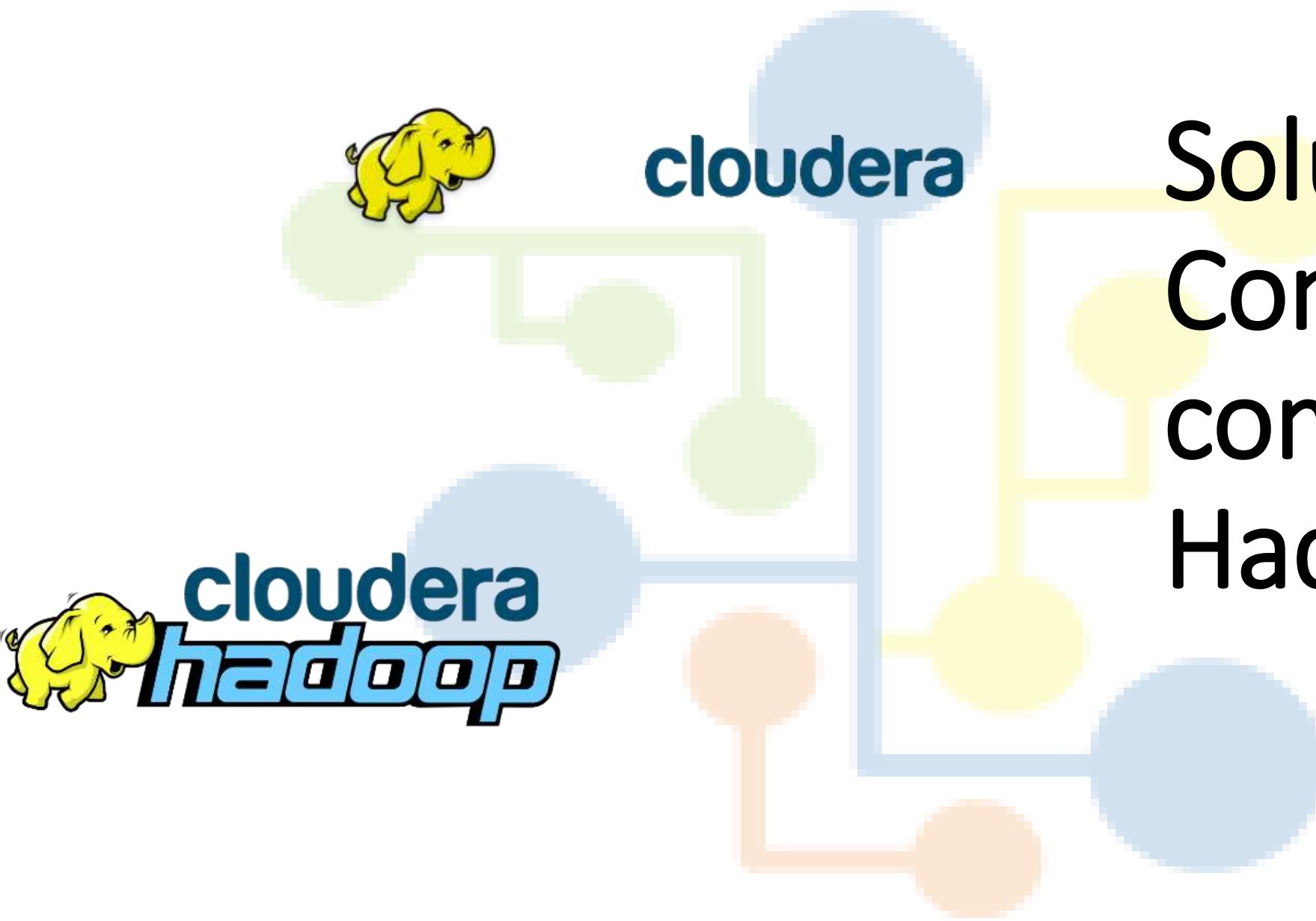
Soluções Comerciais com Hadoop



Amazon Web Services EMR está entre uma das distribuições comerciais do Hadoop com a maior participação no mercado global

Soluções Comerciais com Hadoop



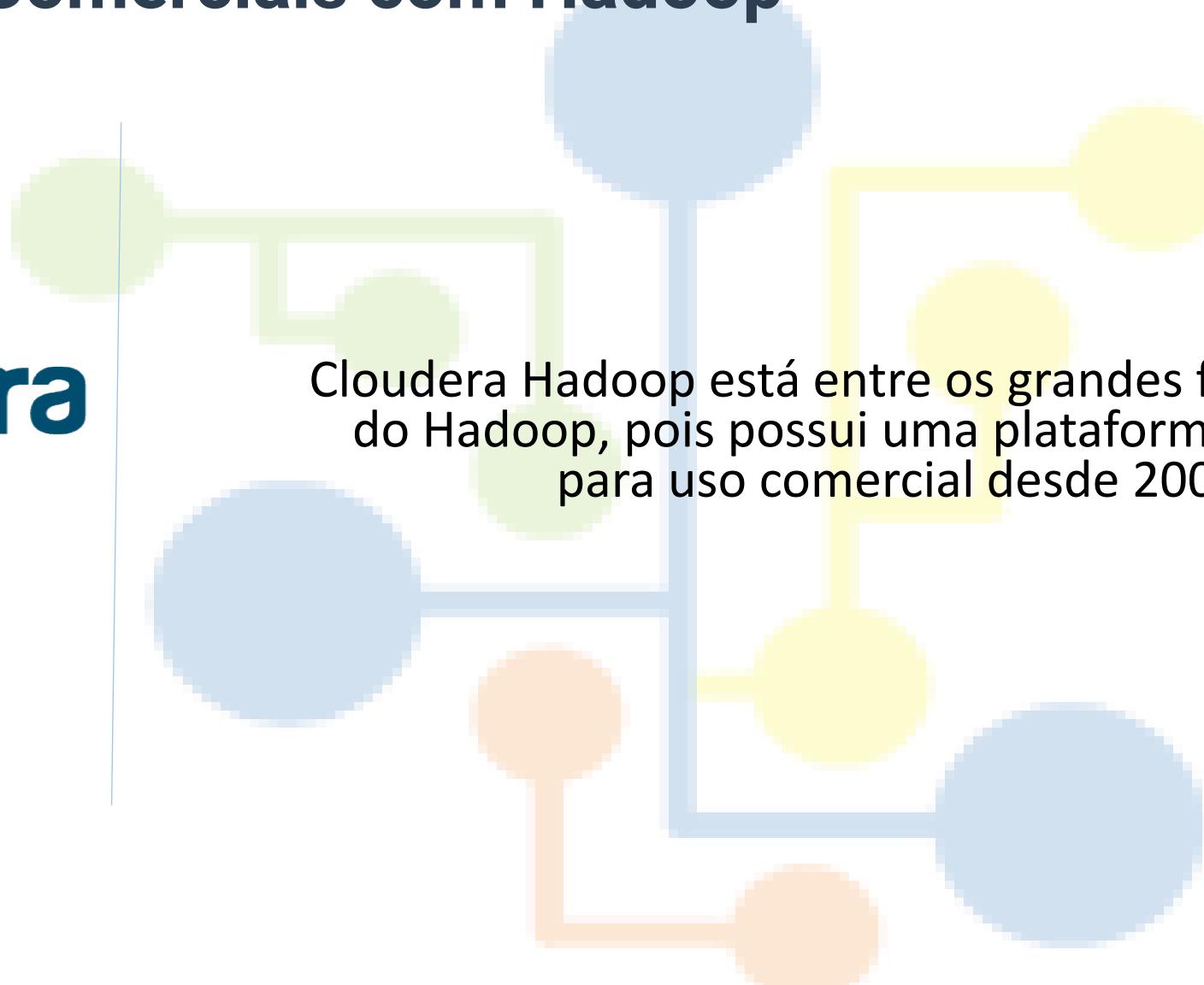


Soluções Comerciais com Hadoop

cloudera
hadoop

Soluções Comerciais com Hadoop

cloudera



Cloudera Hadoop está entre os grandes fornecedores do Hadoop, pois possui uma plataforma confiável para uso comercial desde 2008



Soluções Comerciais com Hadoop

cloudera

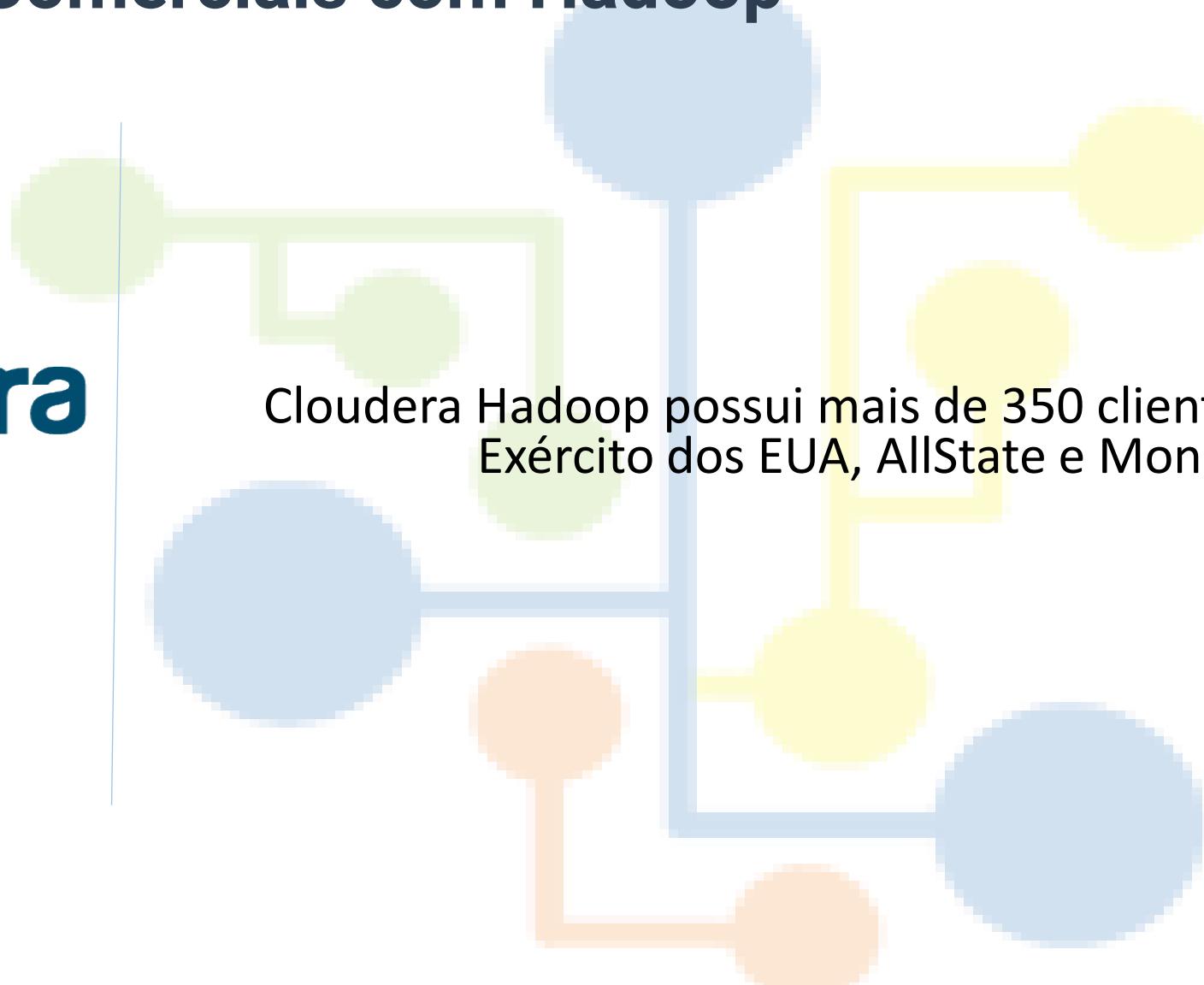


Cloudera, fundada por um grupo de engenheiros do Yahoo, Google e Facebook, está focada em fornecer soluções empresariais do Hadoop



Soluções Comerciais com Hadoop

cloudera

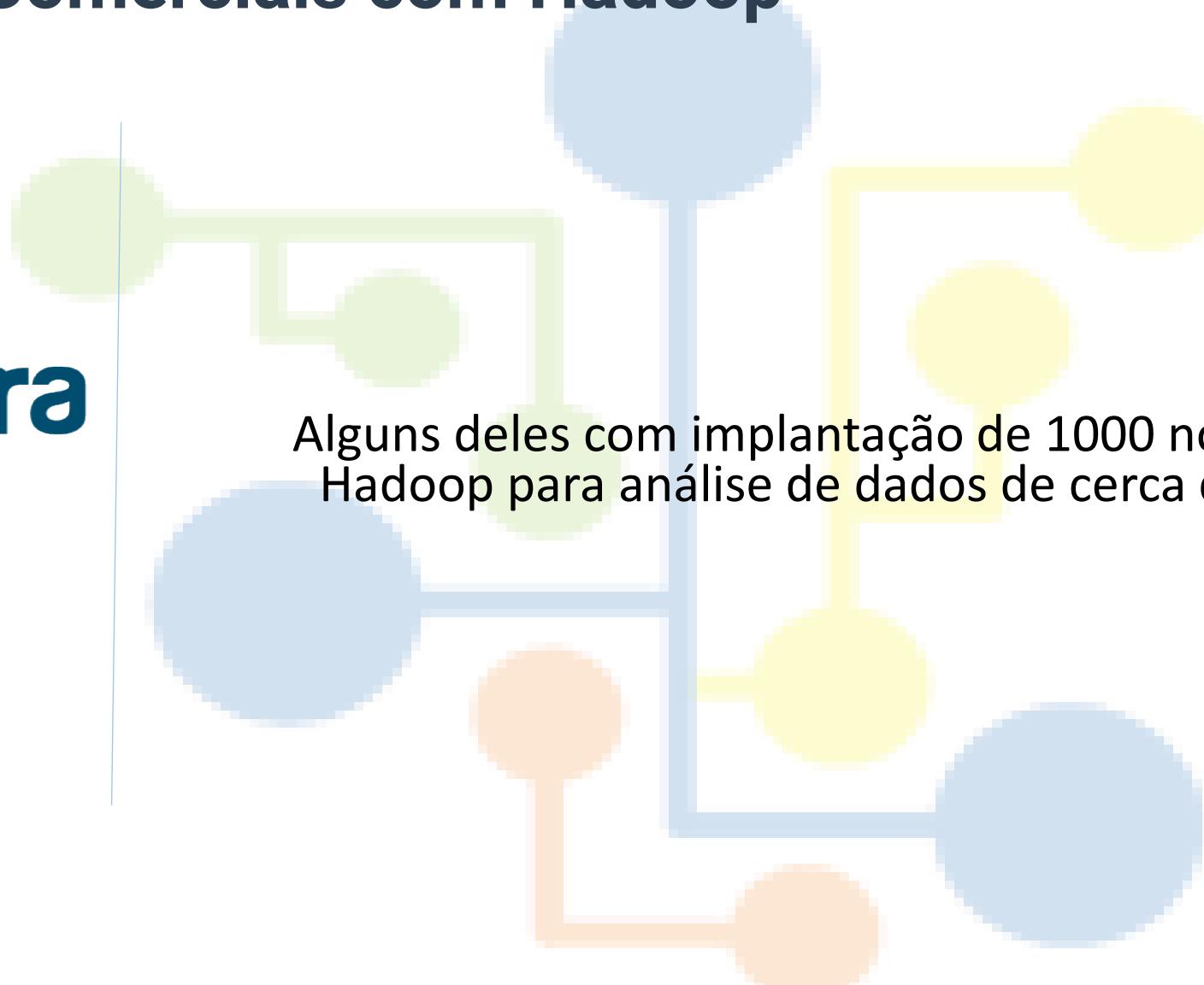


Cloudera Hadoop possui mais de 350 clientes, incluindo o Exército dos EUA, AllState e Monsanto



Soluções Comerciais com Hadoop

cloudera



Alguns deles com implantação de 1000 nós em um cluster Hadoop para análise de dados de cerca de um Petabyte



Soluções Comerciais com Hadoop

cloudera

Cloudera utiliza produtos 100% open-source



Soluções Comerciais com Hadoop

cloudera

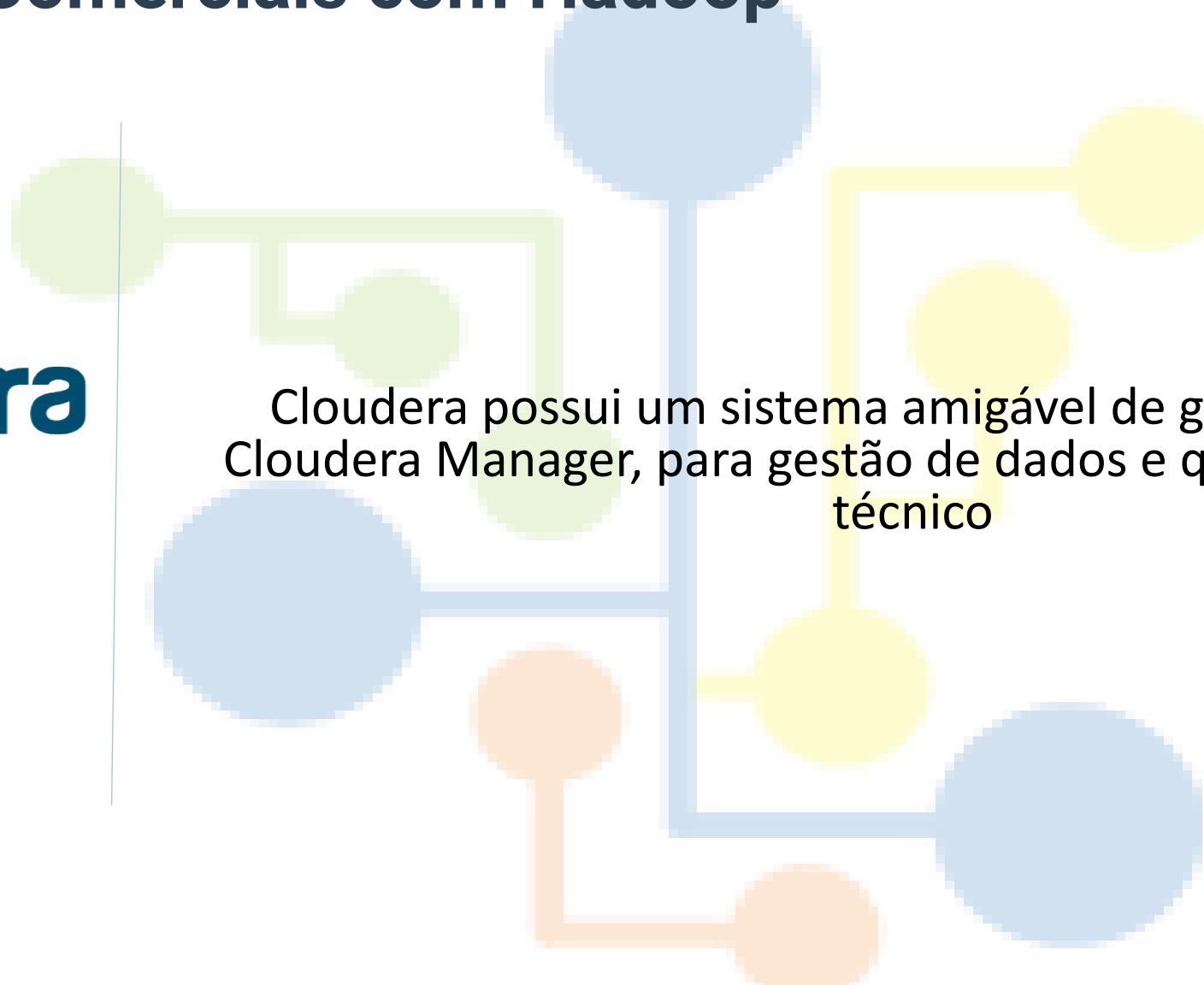
Cloudera utiliza produtos 100% open-source

- Apache Hadoop
- Apache Pig
- Apache Hive
- Apache HBase
- Apache Sqoop



Soluções Comerciais com Hadoop

cloudera



Cloudera possui um sistema amigável de gestão, chamado Cloudera Manager, para gestão de dados e que possui suporte técnico



Soluções Comerciais com Hadoop

cloudera

CDH

BATCH
PROCESSING
(MapReduce,
Hive, Pig)

ANALYTIC
SQL
(Impala)

SEARCH
ENGINE
(Cloudera Search)

MACHINE
LEARNING
(Spark, MapReduce,
Mahout)

STREAM
PROCESSING
(Spark)

3RD PARTY
APPS
(Partners)

WORKLOAD MANAGEMENT (YARN)

STORAGE FOR ANY TYPE OF DATA
UNIFIED, ELASTIC, RESILIENT, SECURE (Sentry)

Filesystem
(HDFS)

Online NoSQL
(HBase)

DATA INTEGRATION (Sqoop, Flume, NFS)

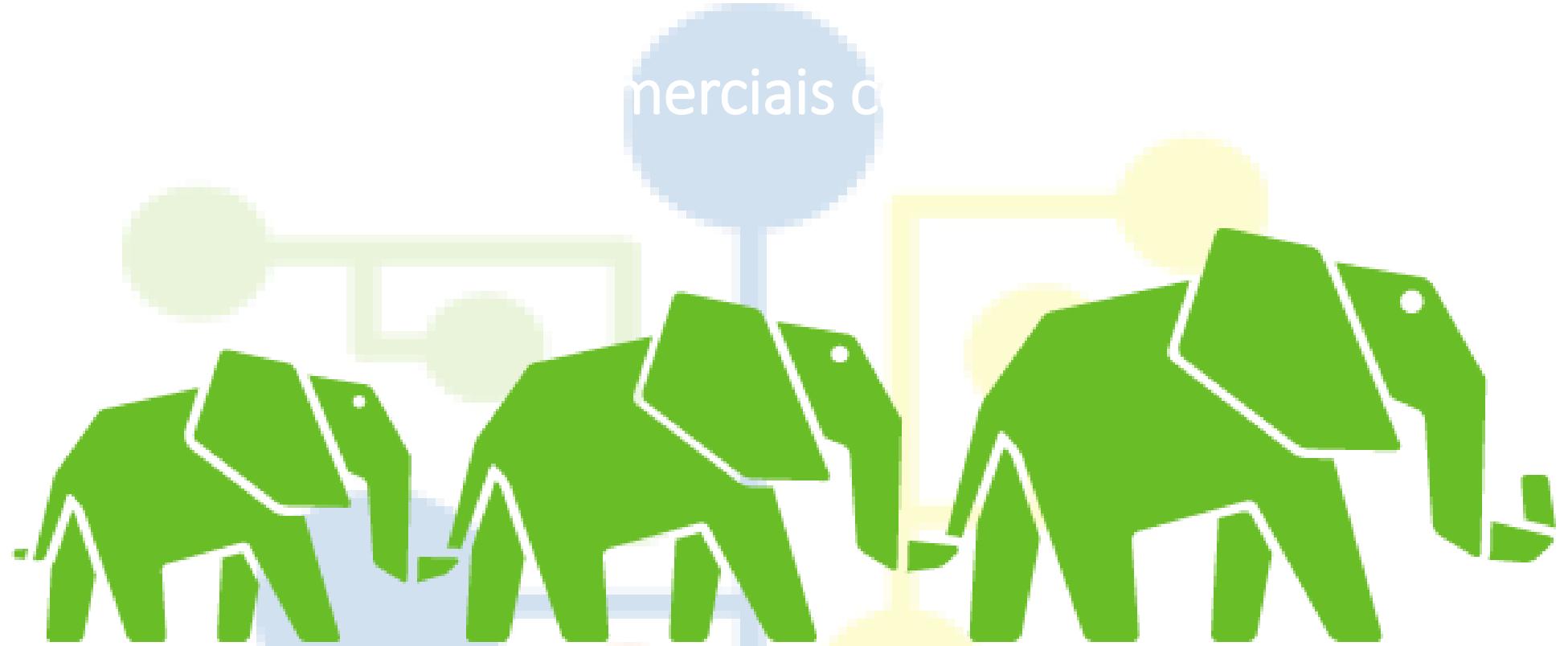


Soluções Comerciais com Hadoop

cloudera

<http://www.cloudera.com>





Hortonworks

Soluções Comerciais com Hadoop



Hortonworks Data Platform (HDP) é uma suite de funcionalidades essenciais para implementação do Hadoop, que pode ser usado para qualquer plataforma tecnológica de dados

Soluções Comerciais com Hadoop



O principal objetivo da Hortonworks é conduzir todas as suas inovações através da plataforma Hadoop e construir um ecossistema de parceiros que acelere o processo de adoção do Hadoop entre as empresas

Soluções Comerciais com Hadoop



Apache Ambari é um exemplo de console de gerenciamento de cluster Hadoop desenvolvido pelo fornecedor Hortonworks para a gestão e monitoramento de clusters Hadoop

Soluções Comerciais com Hadoop



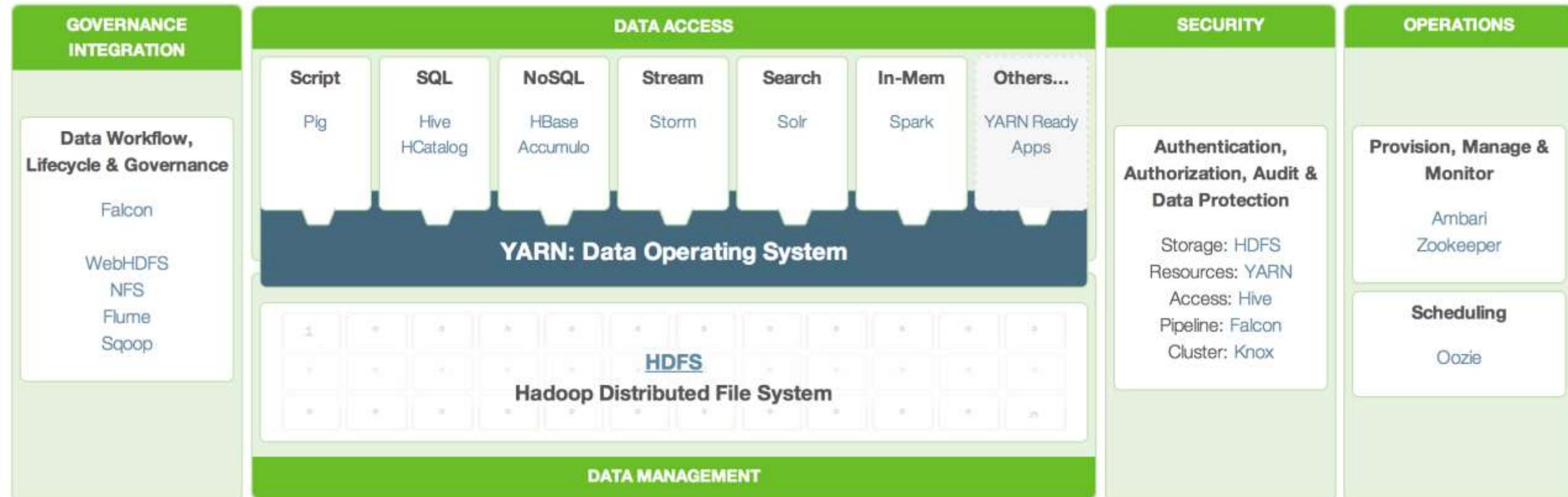
A Hortonworks Hadoop tem atraído mais de 60 novos clientes a cada trimestre com algumas contas gigantes como Samsung, Spotify, Bloomberg e eBay

Soluções Comerciais com Hadoop



A Hortonworks tem atraído fortes parcerias de engenharia com RedHat, Microsoft, SAP e Teradata

Soluções Comerciais com Hadoop



Download available on <http://hortonworks.com/hdp/downloads/>

Soluções Comerciais com Hadoop



Soluções Comerciais com Hadoop



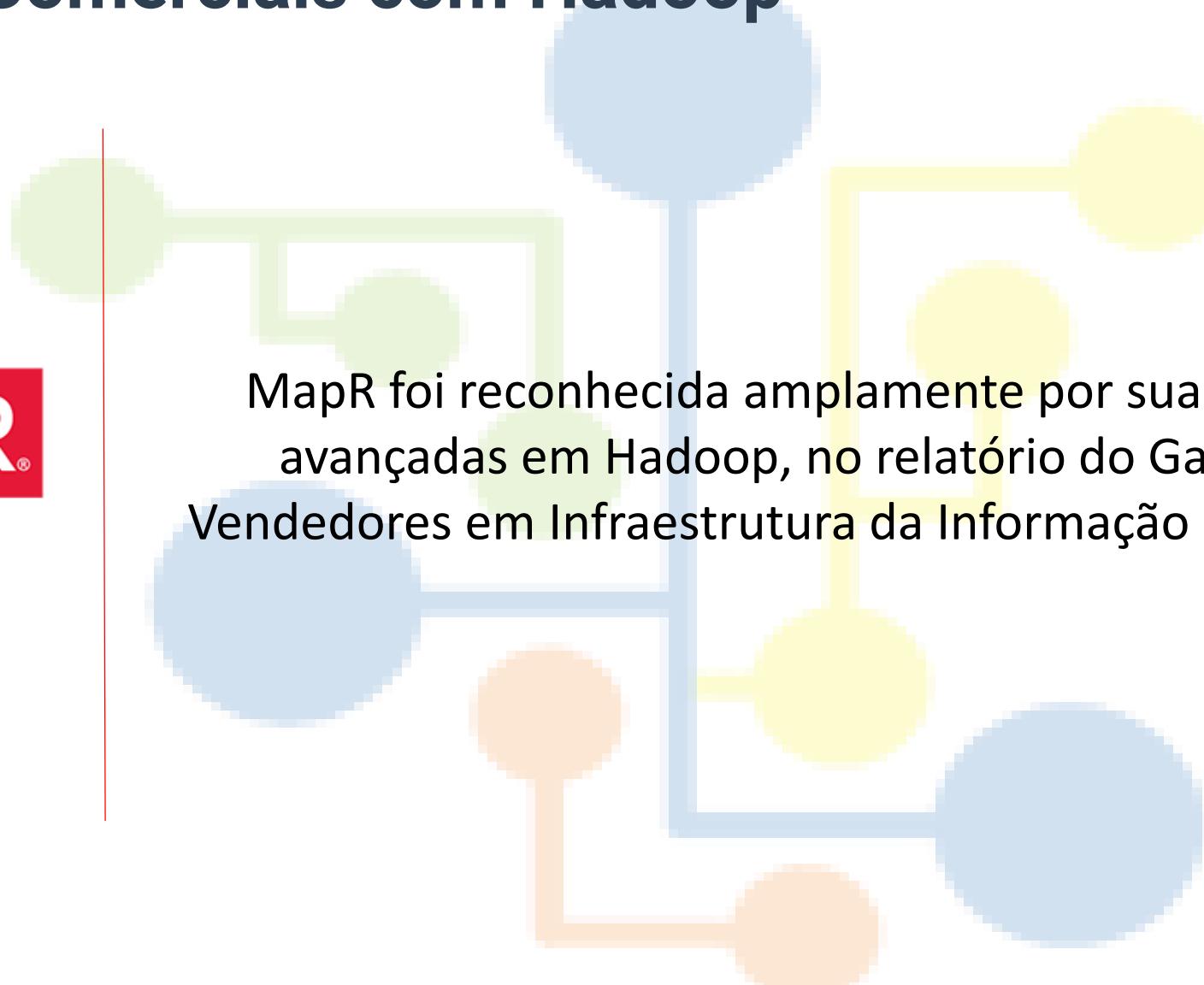
Soluções Comerciais com Hadoop



MapR Data Platform suporta mais de 20 projetos open-source



Soluções Comerciais com Hadoop



An abstract diagram composed of various colored circles (blue, yellow, green, orange) connected by thin lines, forming a complex network or cloud-like structure. It serves as a visual metaphor for data connectivity and infrastructure. A vertical red line is positioned to the left of the text block, aligning with the left edge of the network diagram.

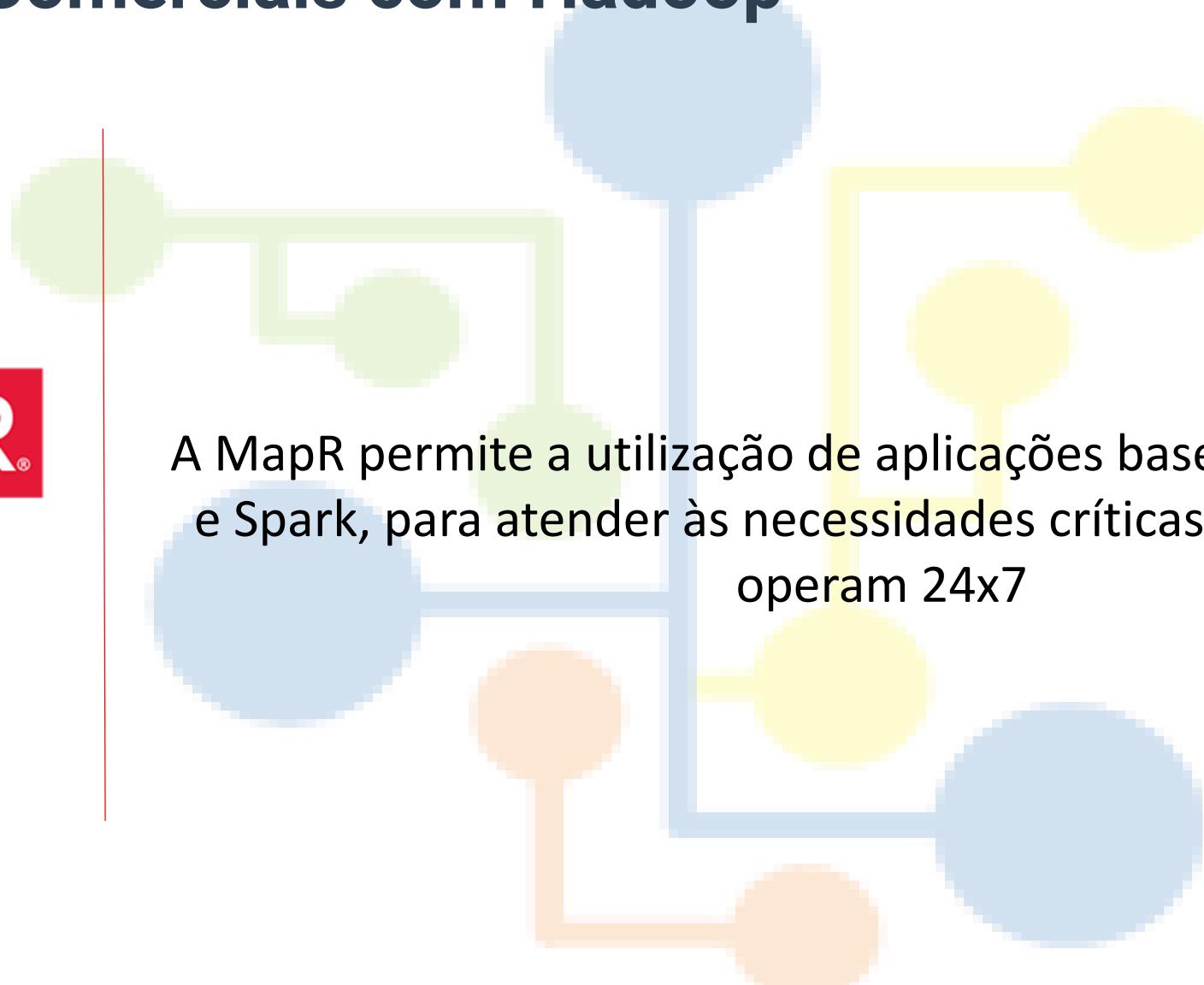
MapR foi reconhecida amplamente por suas distribuições avançadas em Hadoop, no relatório do Gartner "Super Vendedores em Infraestrutura da Informação e Big Data, 2012"

Soluções Comerciais com Hadoop



MapR foi projetada tendo em mente as operações de TI
em Data Centers

Soluções Comerciais com Hadoop



A complex network diagram composed of various colored circles (blue, yellow, green, orange) connected by lines, forming a mesh-like structure. It is positioned on the right side of the slide, partially overlapping the text area. A vertical red line is located to the left of the text block.

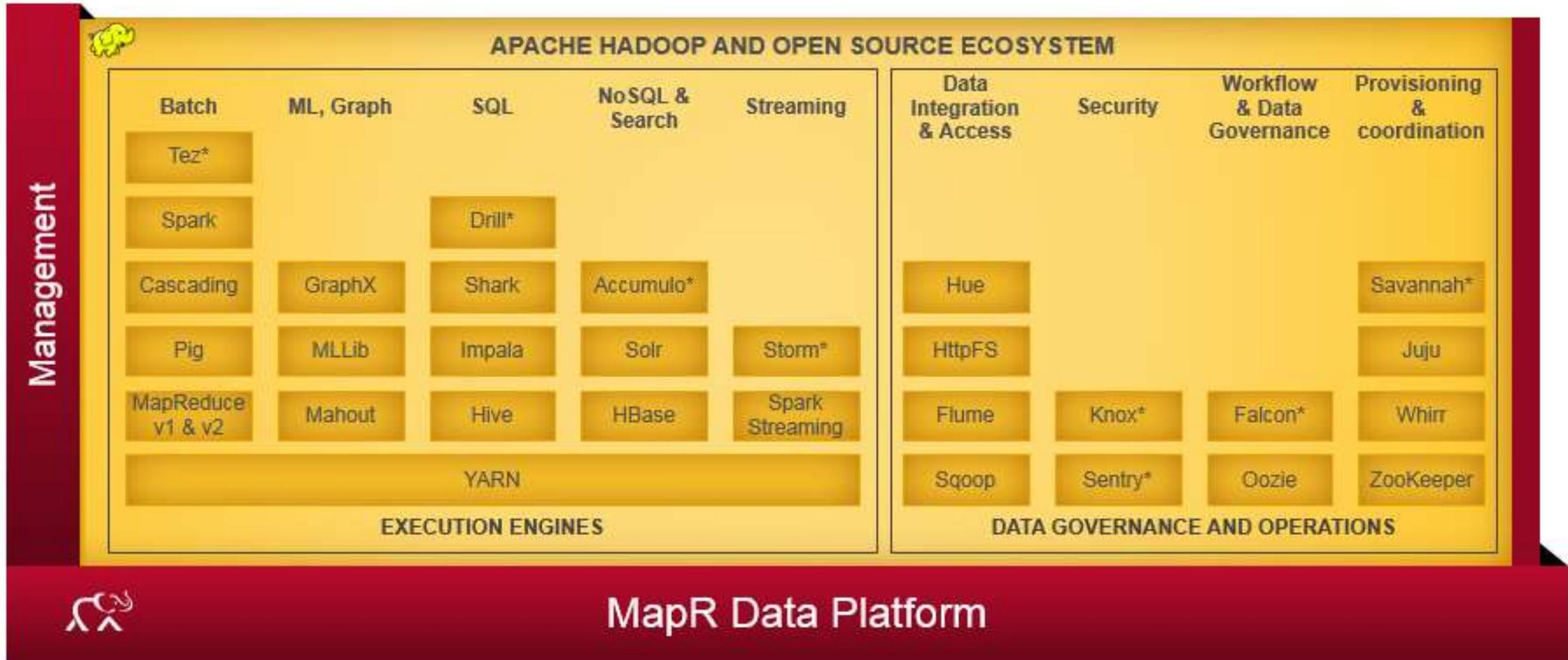
A MapR permite a utilização de aplicações baseadas em Hadoop e Spark, para atender às necessidades críticas de negócio, que operam 24x7

Soluções Comerciais com Hadoop



A MapR suporta amplamente processamento de dados em batch ou streaming de dados em tempo real

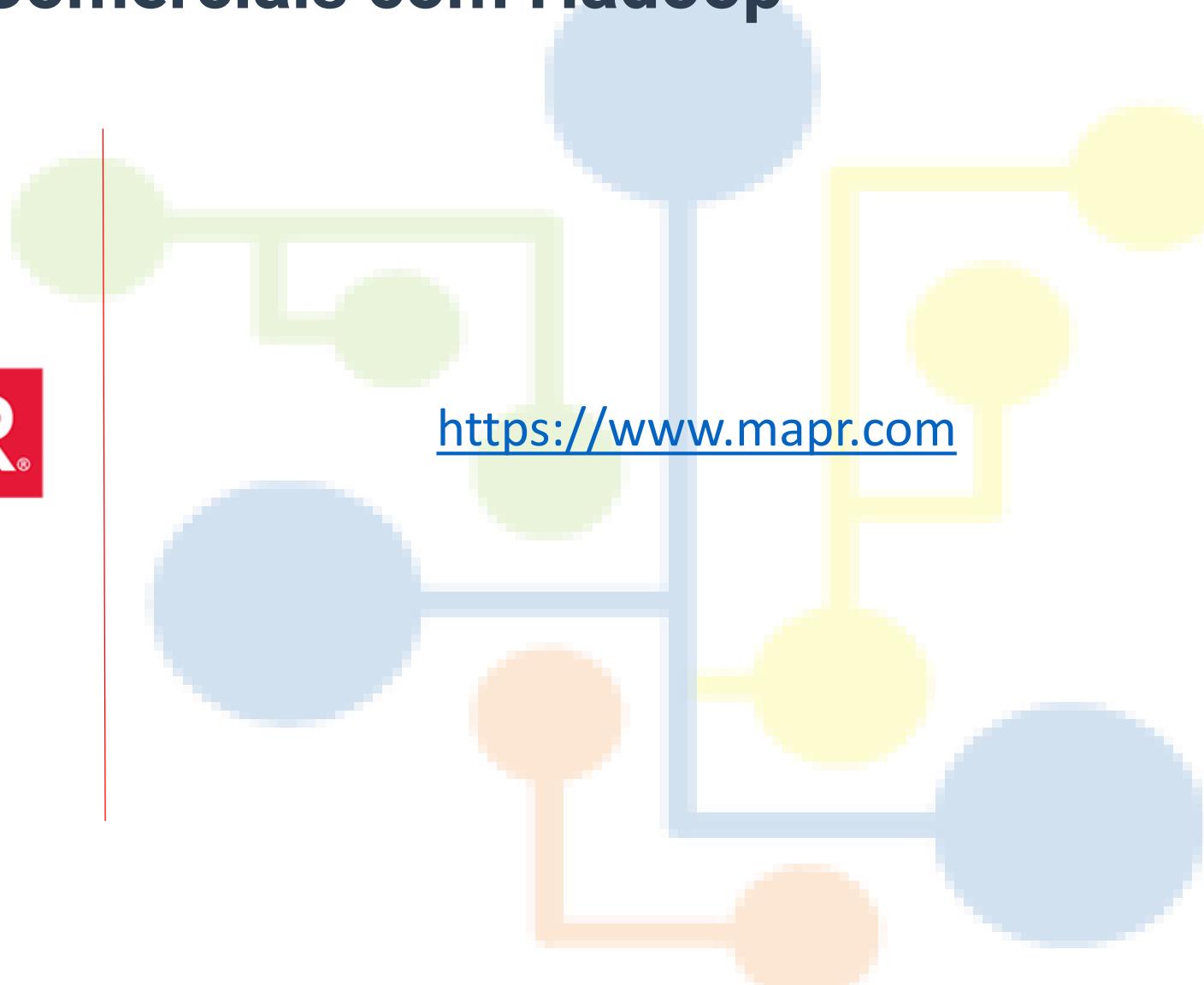
Soluções Comerciais com Hadoop



MapR Data Platform

Download available on: <https://www.mapr.com/products/hadoop-download>

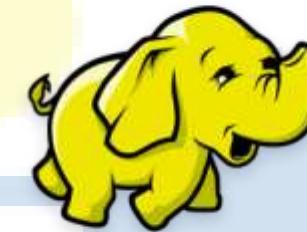
Soluções Comerciais com Hadoop



Soluções Comerciais com Hadoop

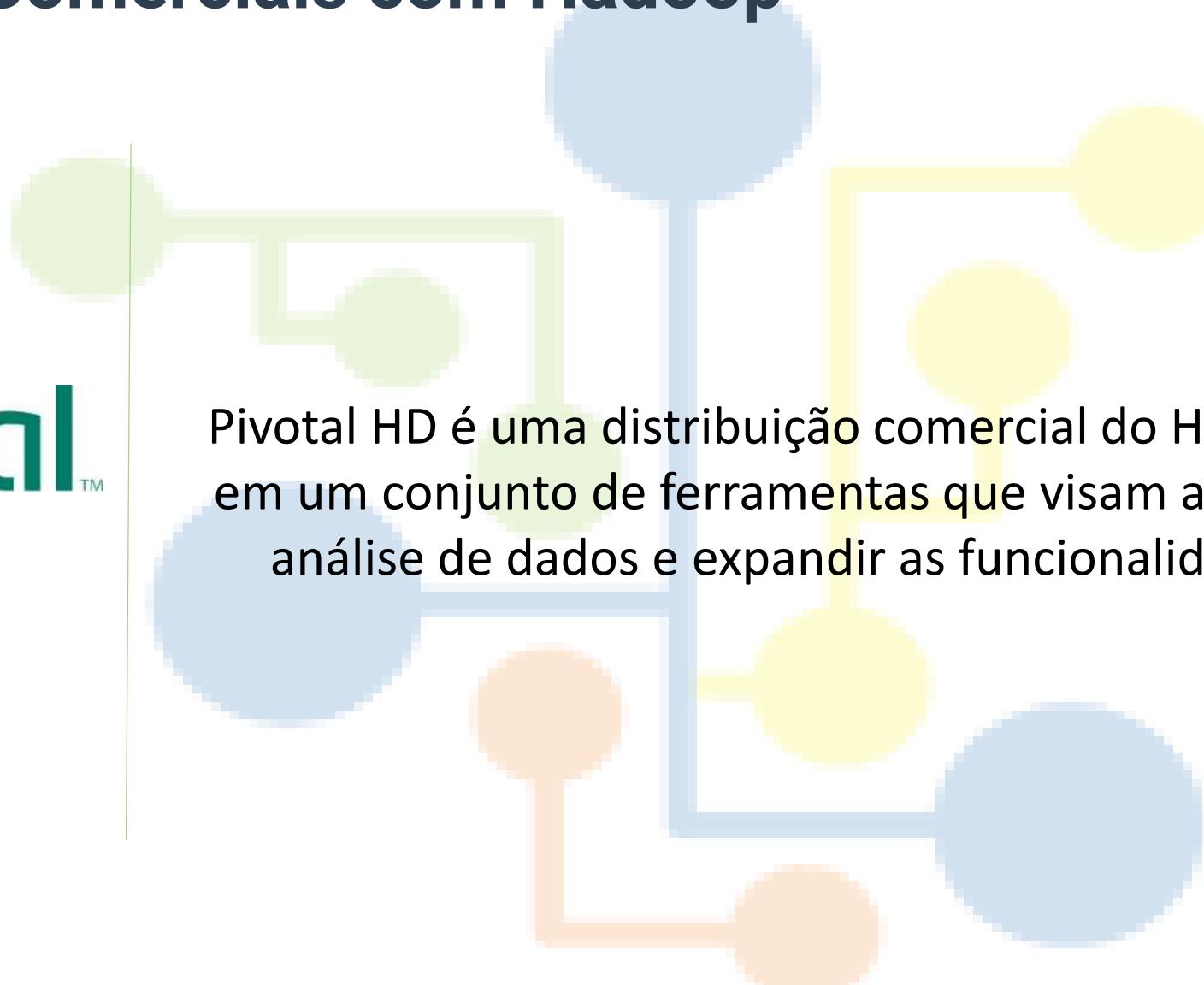
Distribuição
Comercial
Hadoop

PivotalTM



Soluções Comerciais com Hadoop

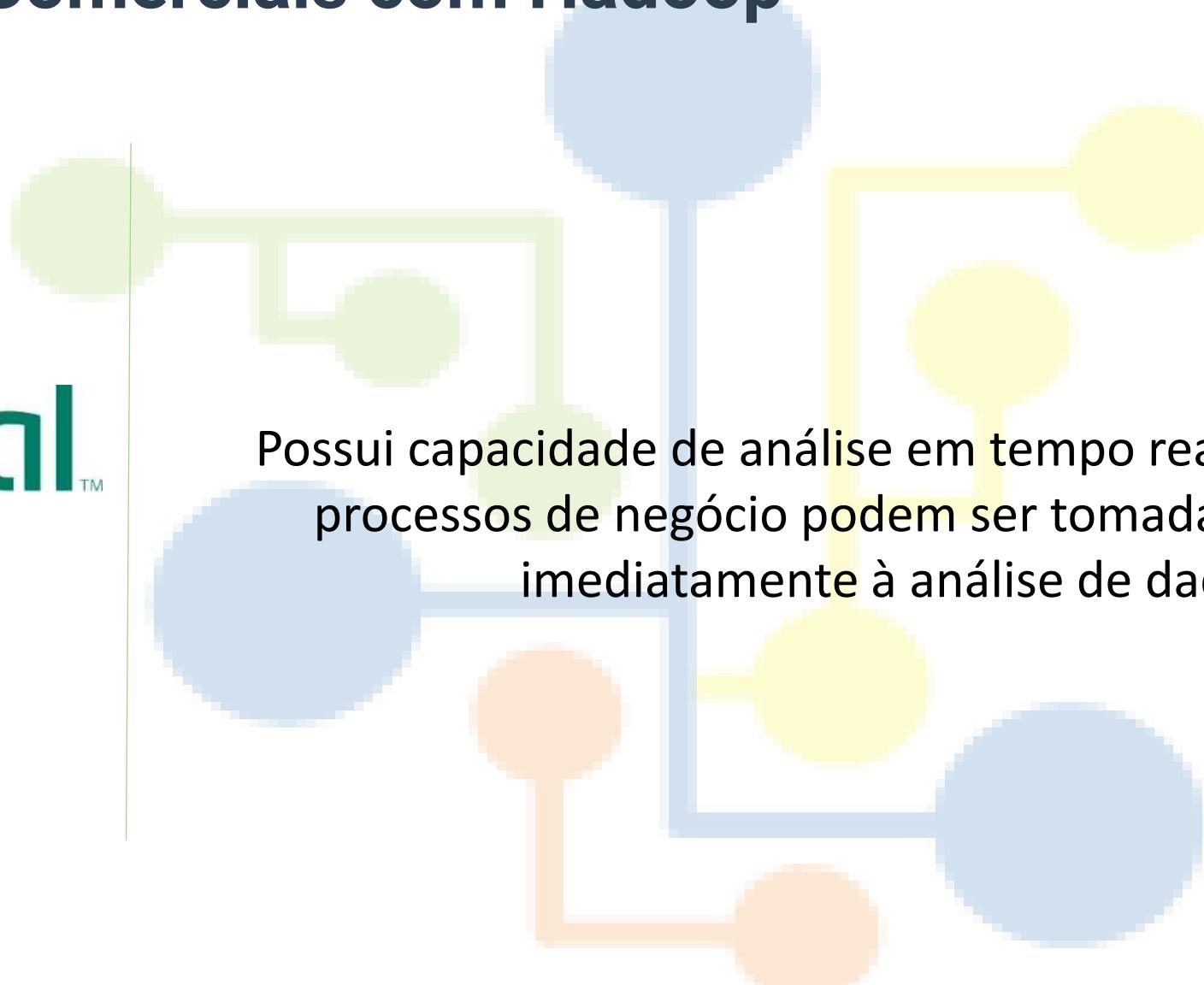
Pivotal™



Pivotal HD é uma distribuição comercial do Hadoop. Ele consiste em um conjunto de ferramentas que visam acelerar projetos de análise de dados e expandir as funcionalidades do Hadoop

Soluções Comerciais com Hadoop

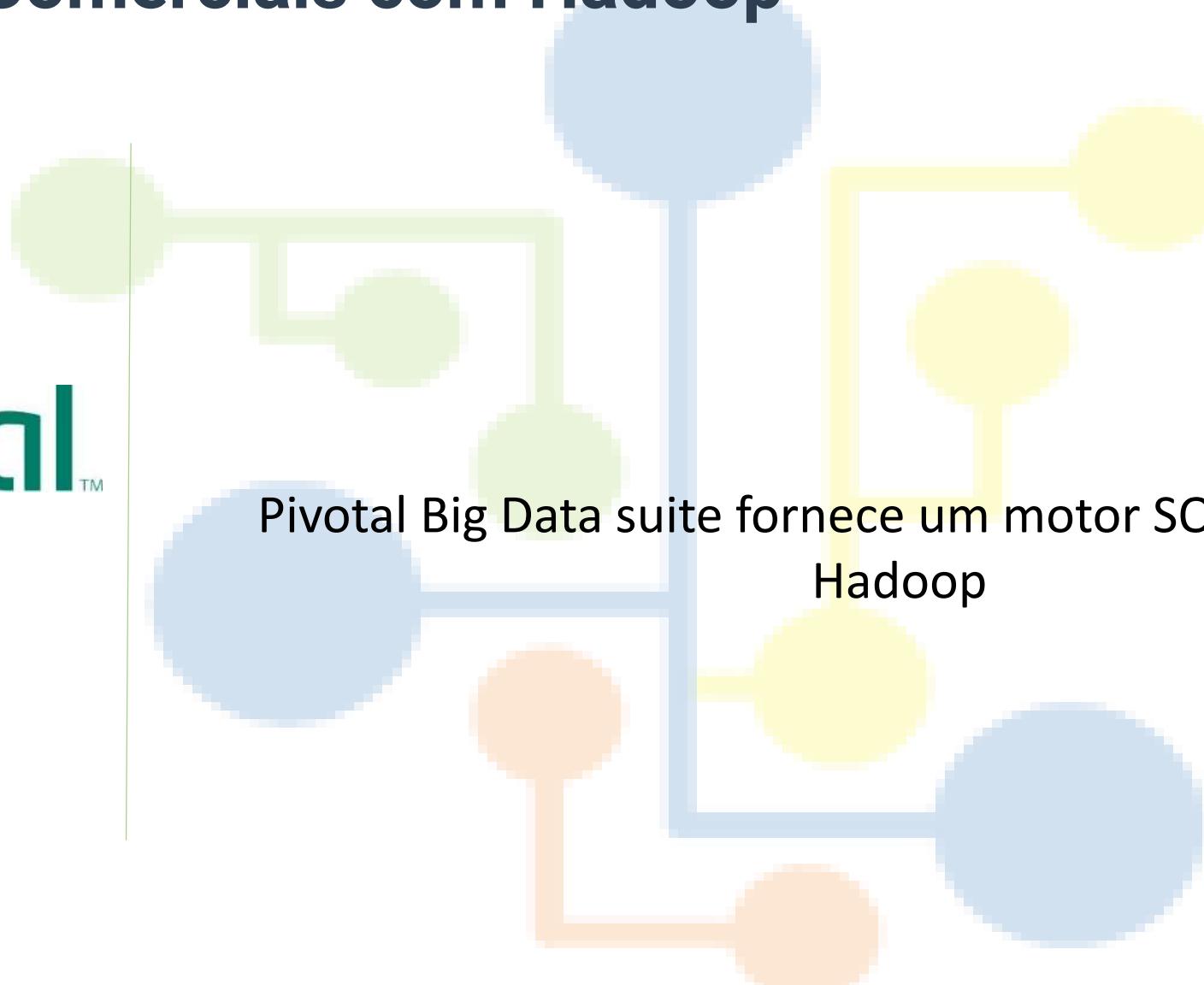
Pivotal™



Possui capacidade de análise em tempo real e decisões de processos de negócio podem ser tomadas quase que imediatamente à análise de dados

Soluções Comerciais com Hadoop

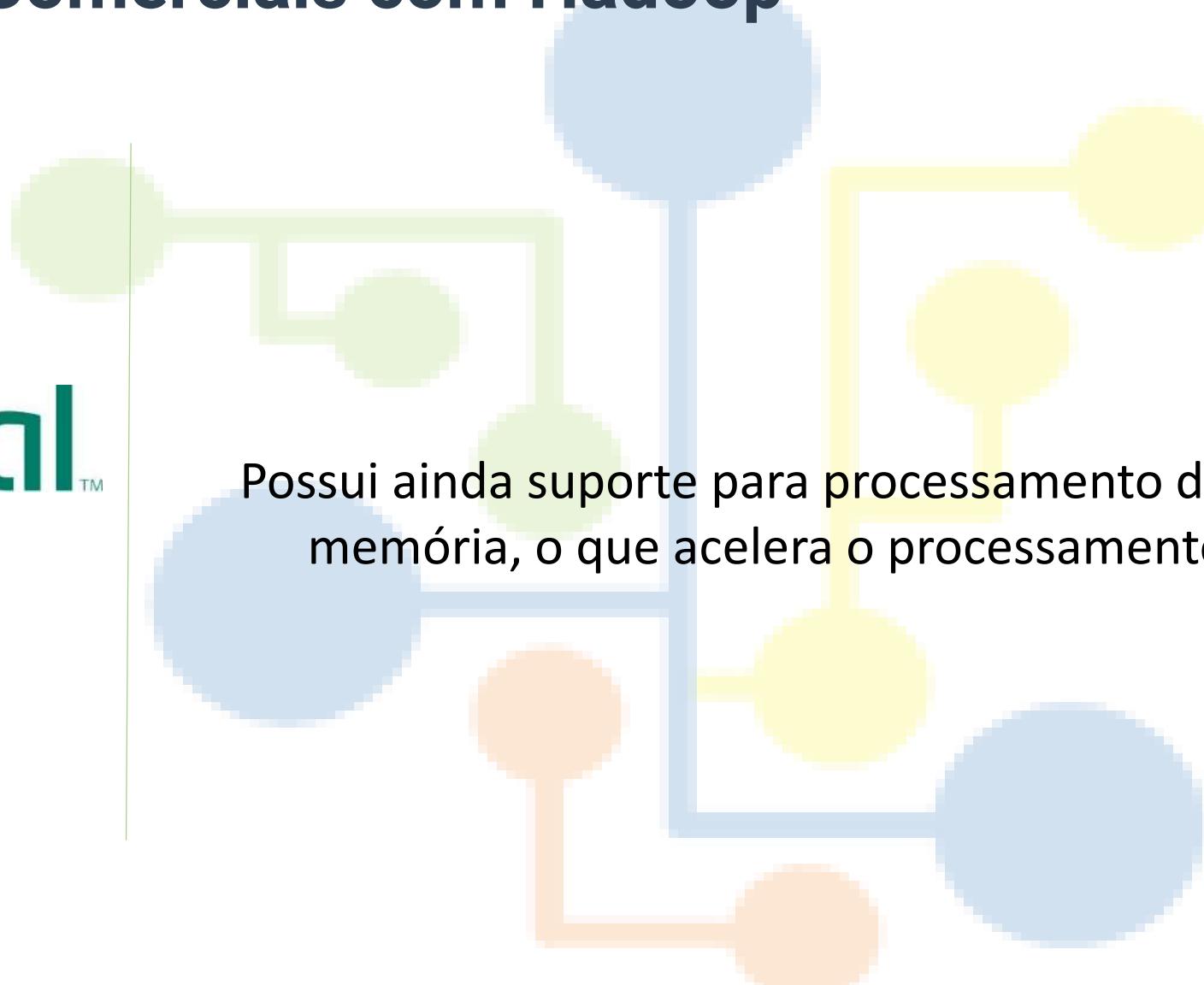
Pivotal™



Pivotal Big Data suite fornece um motor SQL nativo para o
Hadoop

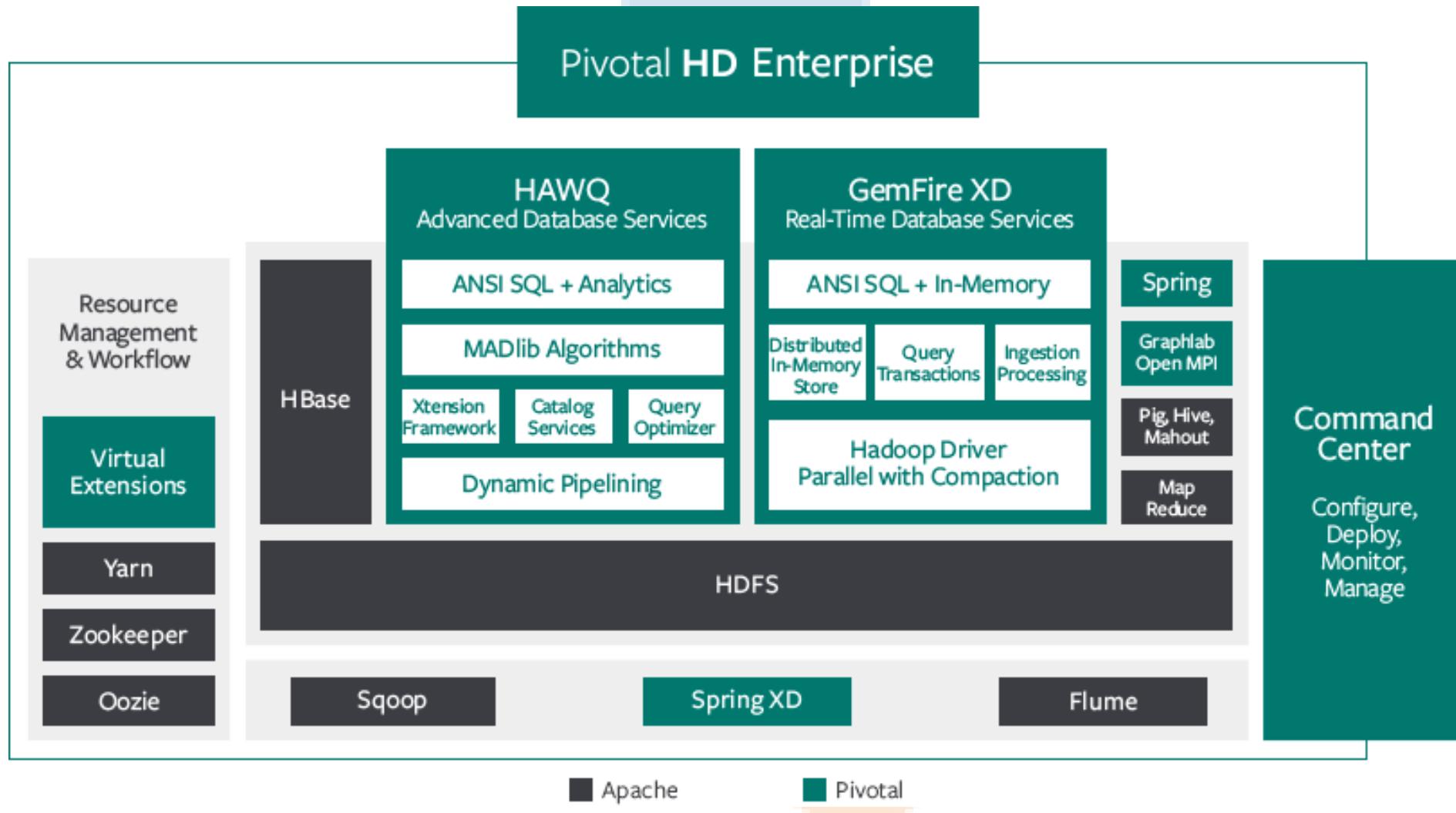
Soluções Comerciais com Hadoop

Pivotal™



Possui ainda suporte para processamento de Big Data em memória, o que acelera o processamento de dados

Soluções Comerciais com Hadoop

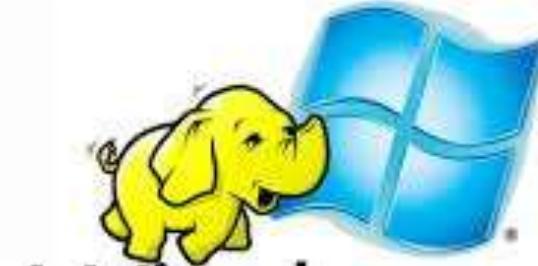


Soluções Comerciais com Hadoop

Pivotal™

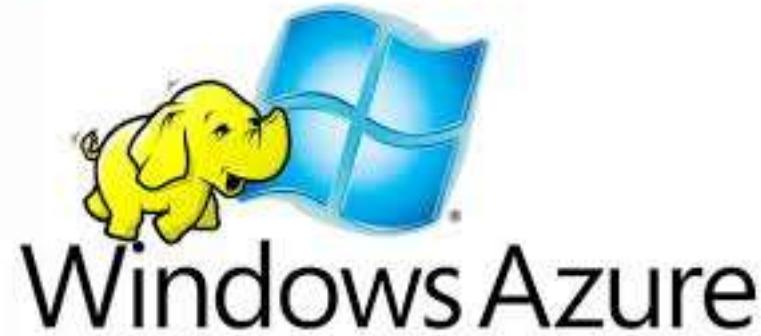
<http://pivotal.io>

Soluções Comerciais com Hadoop



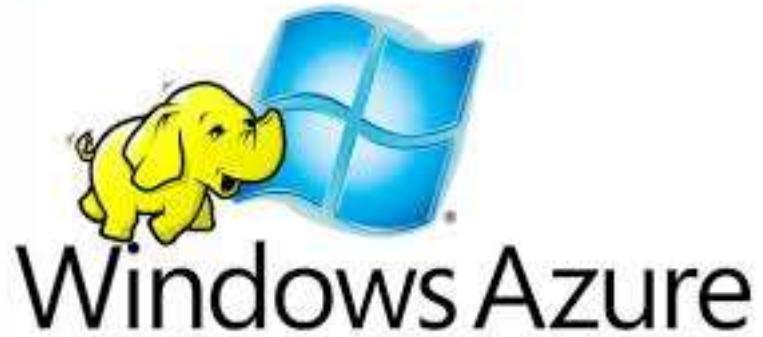
Windows Azure

Soluções Comerciais com Hadoop



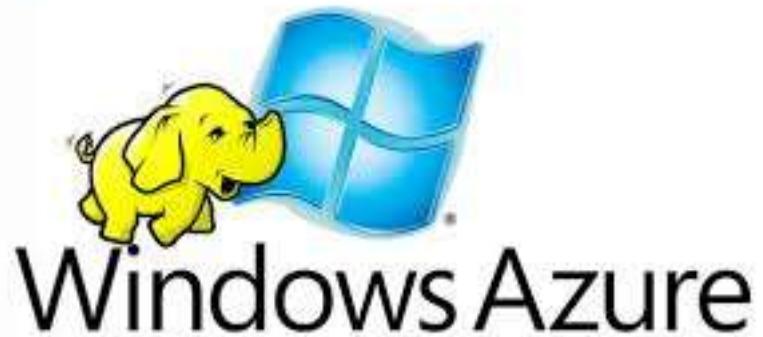
Azure HDInsight é uma distribuição Apache Hadoop distribuída em Cloud

Soluções Comerciais com Hadoop



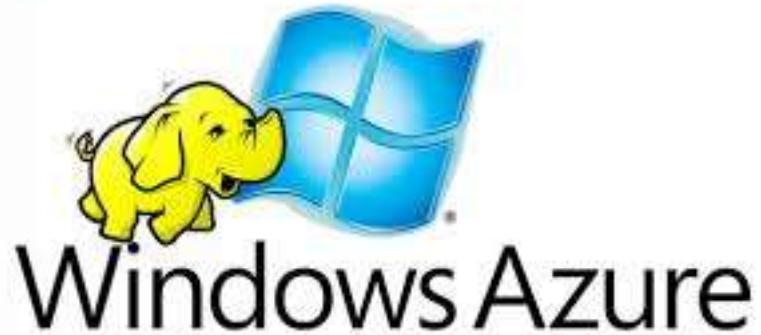
O Azure HDInsight consegue lidar com quantidades de dados, de terabytes até petabytes, permitindo a inclusão de nodes sob demanda

Soluções Comerciais com Hadoop



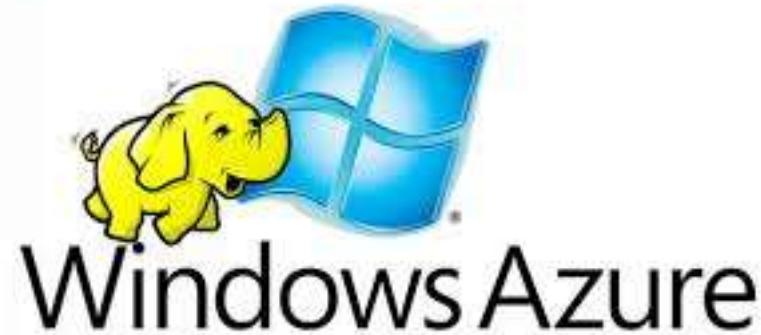
Por ser 100% Apache Hadoop, o HDInsight pode processar dados semi-estruturados ou não-estruturados, tais como clicks em páginas web, posts em mídia social, logs de servidores, dados de sensores, etc...

Soluções Comerciais com Hadoop



O HDInsight também possui extensões para programação em C#, Java e .NET, que podem ser usadas para criar, configurar, submeter e monitorar jobs Hadoop

Soluções Comerciais com Hadoop

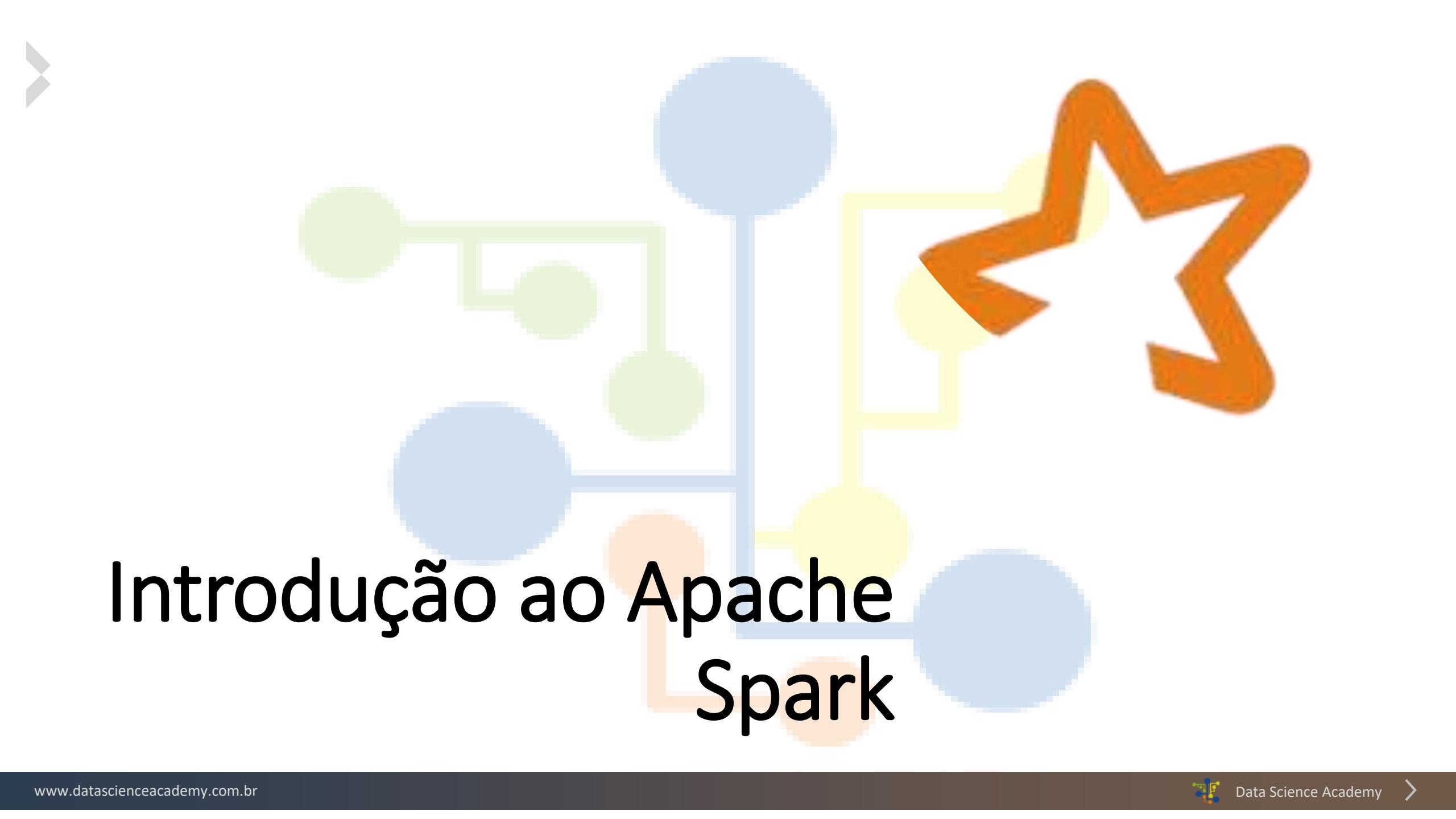


Por ser integrado com Excel®, o HDInsight permite visualizar e analisar dados do Hadoop, de forma que seja familiar aos usuários finais

Soluções Comerciais com Hadoop



<https://azure.microsoft.com/en-us/services/hdinsight>



Introdução ao Apache Spark

Introdução ao Apache Spark



Introdução ao Apache Spark



<http://spark.apache.org>

Apache Spark é um engine rápido e de uso geral para processamento de dados em larga escala

Introdução ao Apache Spark



<http://spark.apache.org>

É显著mente mais rápido que o Hadoop MapReduce e vem ganhando popularidade

Introdução ao Apache Spark



<http://spark.apache.org>

Utiliza o Hadoop (HDFS) como base, mas pode ser usado com Cassandra, HBase e MongoDB

Introdução ao Apache Spark



<http://spark.apache.org>

Pode ser usado com linguagens
Python, R, Scala e Java

Introdução ao Apache Spark



<http://spark.apache.org>

Usado por empresas como Globo.com, Yelp,
Washington Post, Yahoo e Twitter

Introdução ao Apache Spark



| | |
|-----------------------|--|
| Velocidade | Sua velocidade de execução pode ser até 100x mais rápido que o Hadoop MapReduce em memória e 10x em disco |
| Facilidade de uso | Aplicações podem ser escritas em Java, Scala, R e Python |
| Generalidade | Combina SQL Streaming e análise complexa, além do uso de ferramentas de alto nível como Spark SQL, MLlib para Machine Learning, GraphX e Spark Streaming |
| Integração com Hadoop | Executa sobre o YARN cluster manager e permite leitura e escrita de dados no HDFS |

Introdução ao Apache Spark



<http://spark.apache.org>

Spark é um projeto open source, mantido por uma comunidade de desenvolvedores que foi criado em 2009 na Universidade da Califórnia, Berkeley

Introdução ao Apache Spark



<http://spark.apache.org>

Os desenvolvedores estavam trabalhando com [Hadoop MapReduce](#) e perceberam ineficiências na execução de computação iterativa

Introdução ao Apache Spark



<http://spark.apache.org>

Em pouco tempo, Apache Spark tem se tornado o mecanismo de processamento de Big Data para a próxima geração e está sendo aplicado em todo o mercado de dados mais rápido do que nunca

Introdução ao Apache Spark



<http://spark.apache.org>

O Apache Spark oferece basicamente
3 principais benefícios:

Introdução ao Apache Spark



<http://spark.apache.org>

1- Facilidade de uso – é possível desenvolver API's de alto nível em Java, Scala, Python e R, que permitem focar apenas no conteúdo a ser computado, sem se preocupar com configurações de baixo nível e extremamente técnicas.

Introdução ao Apache Spark



<http://spark.apache.org>

2- Velocidade – Spark é veloz, permitindo uso iterativo e processamento rápido de algoritmos complexos. Velocidade é uma característica especialmente importante no processamento de grandes conjuntos de dados e pode fazer a diferença entre analisar os dados de forma interativa ou ficar aguardando vários minutos pelo fim de cada processamento. Com Spark, o processamento é feito em memória.

Introdução ao Apache Spark



<http://spark.apache.org>

3- Uso geral – Spark permite a utilização de diferentes tipos de computação, como processamento de linguagem SQL (SQL Spark), processamento de texto, [Machine Learning](#) (MLlib) e processamento de grafos (GraphX). Estas características fazem do Spark uma excelente opção para projetos de [Big Data](#).

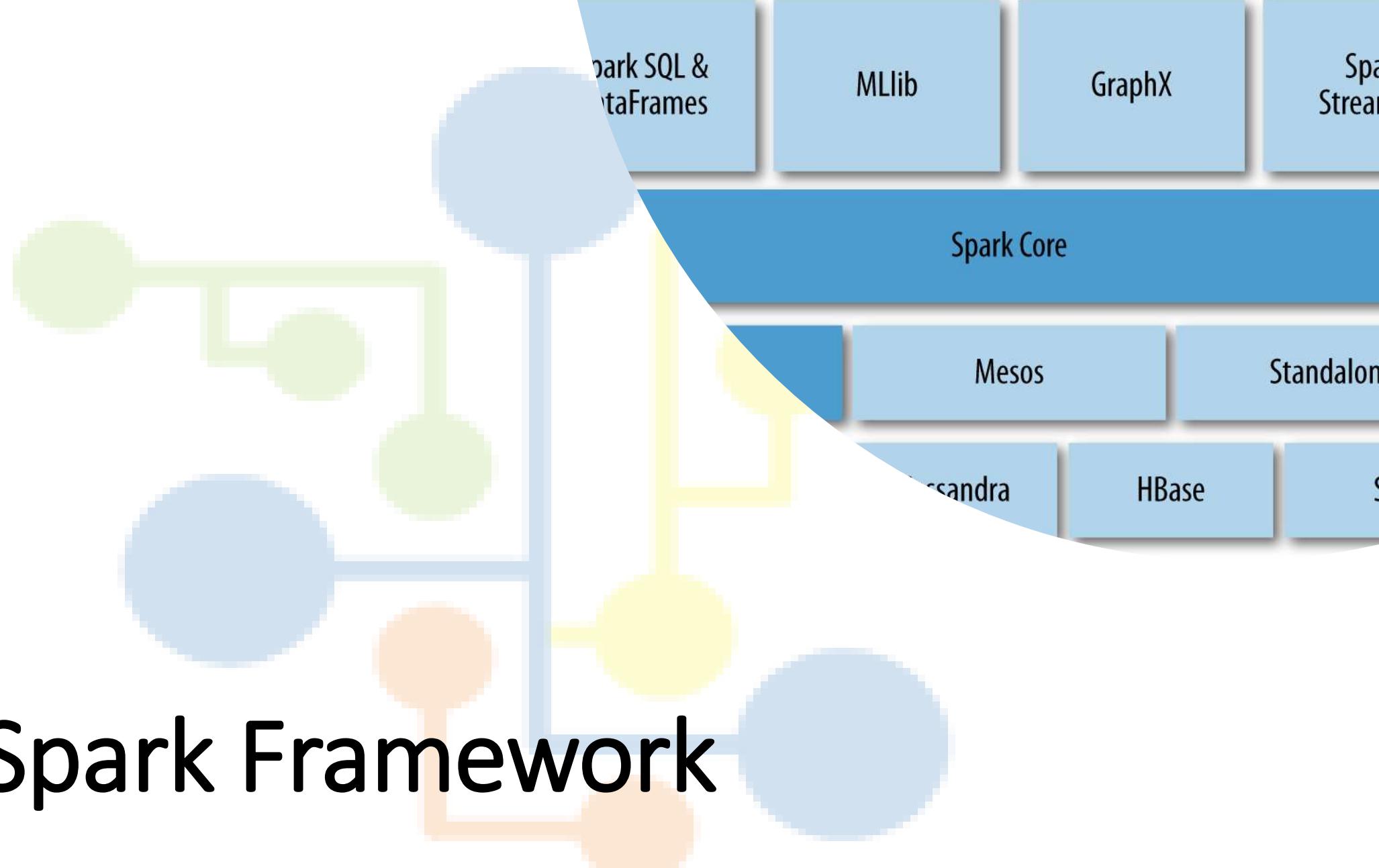
Introdução ao Apache Spark



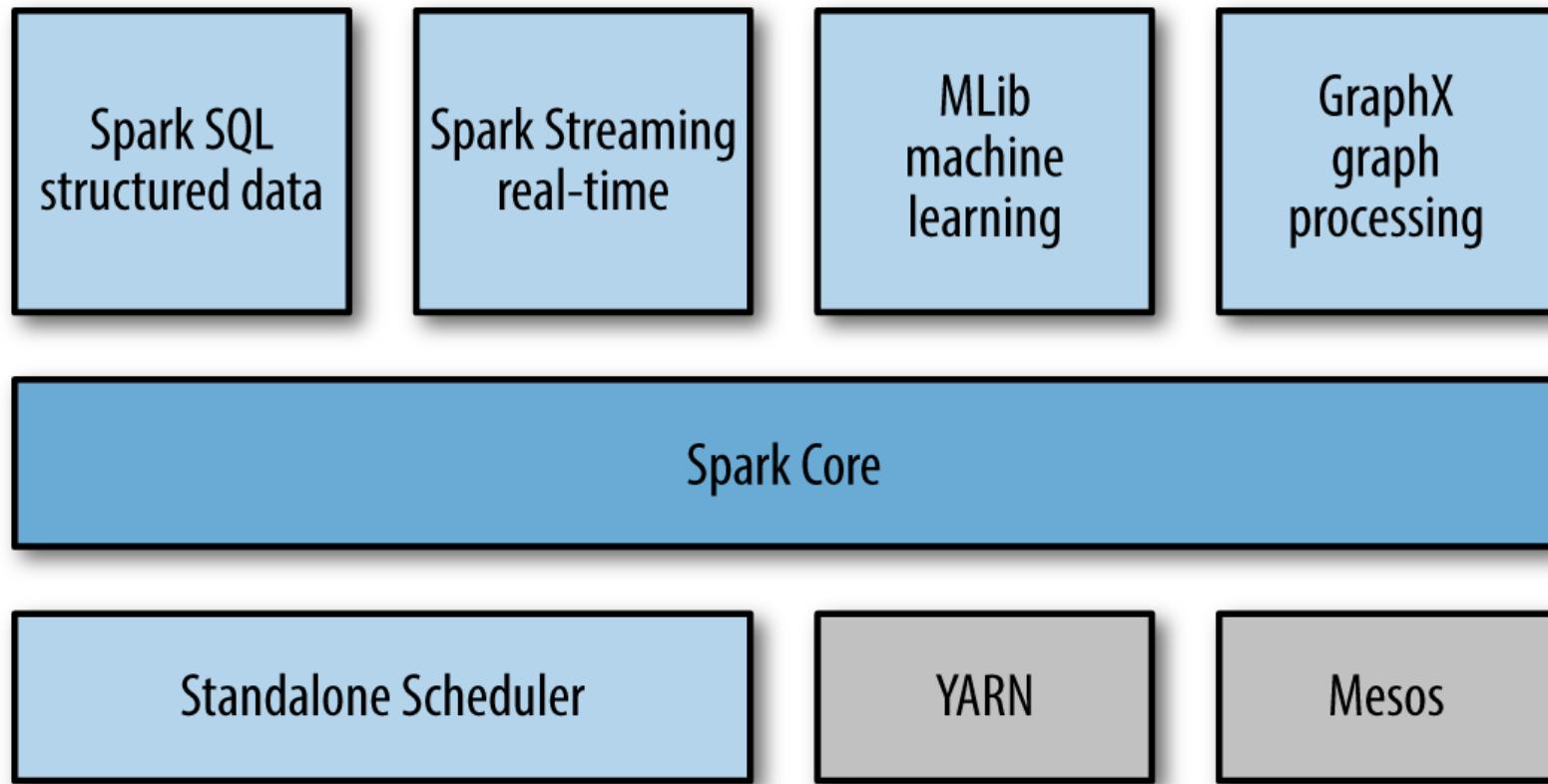
<http://spark.apache.org>

O projeto Spark contém diversos componentes integrados. Basicamente, Spark é um engine de computação, responsável por agendar, distribuir e monitorar aplicações de diversas tarefas de processamento através de diferentes servidores em cluster.

Spark Framework



Introdução ao Apache Spark



Introdução ao Apache Spark

Spark Core



Contém as funcionalidades básicas do Spark, incluindo componentes para agendamento de tarefas, gestão de memória, recuperação de falha e sistemas de armazenamento.

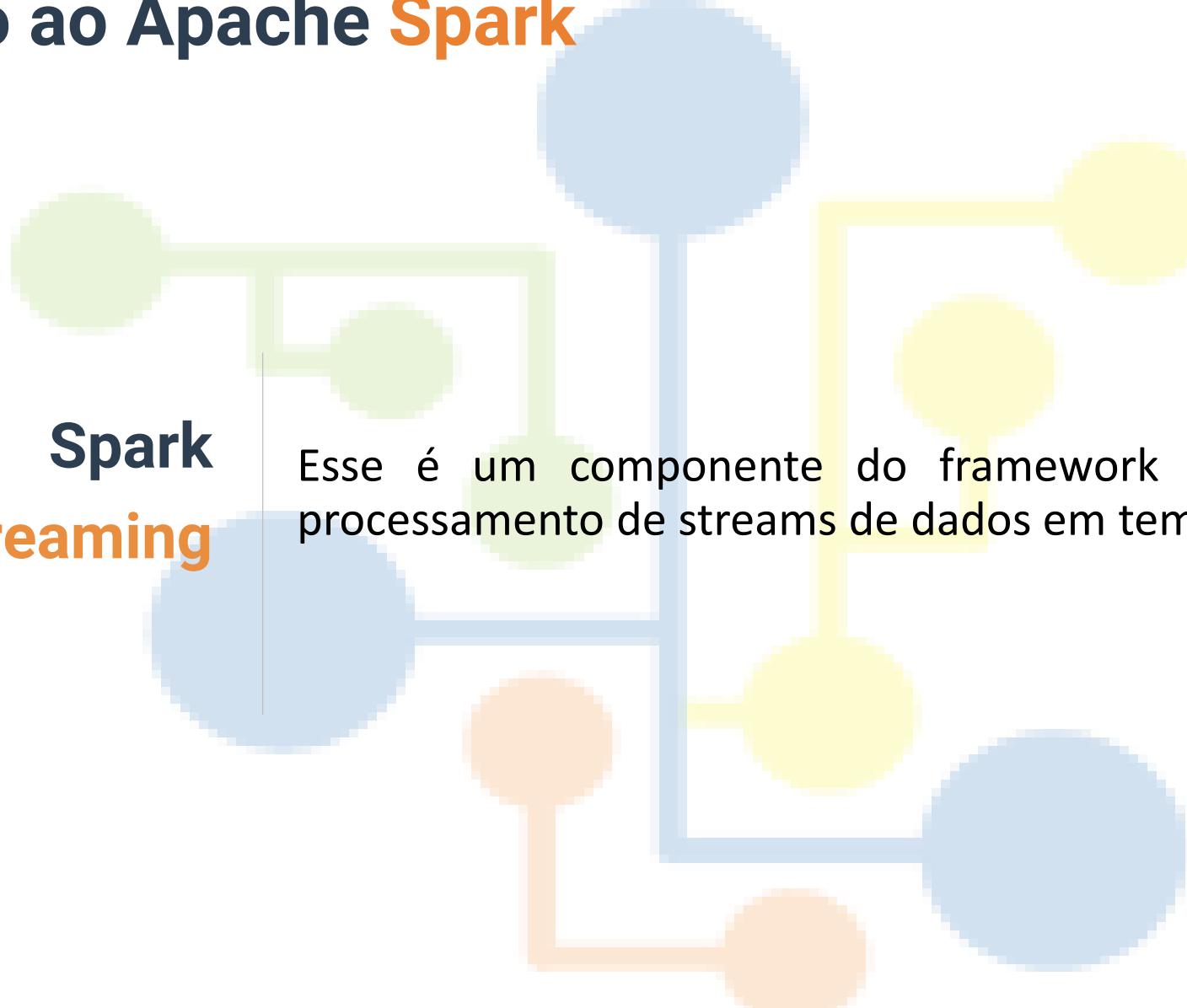
Introdução ao Apache Spark

Spark SQL

Spark SQL é um pacote para tarefas com dados estruturados. Ele permite realizar queries nos dados através de linguagem SQL, além de suportar diversas fontes de dados como [Hive](#) e [JSON](#).

Introdução ao Apache Spark

Spark Streaming



Esse é um componente do framework Spark para processamento de streams de dados em tempo real.

Introdução ao Apache Spark

Spark
MLlib

A biblioteca MLlib é uma funcionalidade para Machine Learning.

Introdução ao Apache Spark

Spark GraphX



O GraphX é um biblioteca para manipulação de grafos e computação em paralelo.

Introdução ao Apache Spark

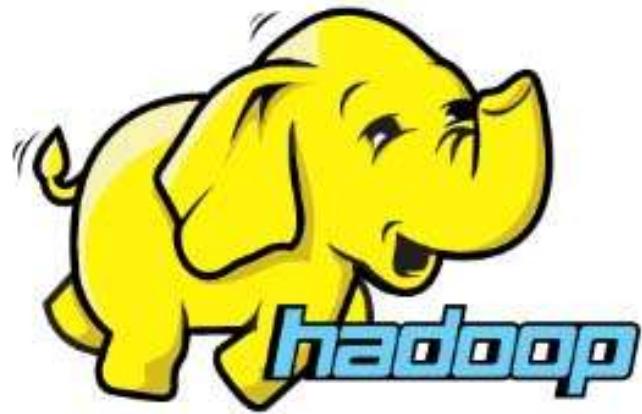


<http://spark.apache.org>

O resultado de um projeto de Big Data, pode ser a criação de um sistema de análise de dados em tempo real, que pode se tornar o componente de uma aplicação de negócio.

Introdução ao Apache Spark

Quando se trata de Hadoop e Spark, duas perguntas são frequentes:



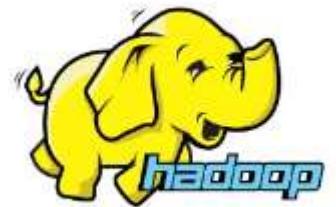
Introdução ao Apache Spark

Quando se trata de Hadoop e Spark, duas perguntas são frequentes:

1- Já estou usando Hadoop, devo usar o Spark?

2- Estou pensando em usar o Hadoop, devo desistir e usar o Spark?

Vamos investigar as diferenças entre Hadoop e Spark e responder estas perguntas!



Introdução ao Apache Spark



O Hadoop é a plataforma original do Big Data, que tem sido usado e testado no mercado. Permite trabalhar com Petabytes de dados, habilitando a análise de quantidades massivas de dados.

O Hadoop possui um ecossistema bem definido que permite estender suas funções, como no caso da utilização do Pig, Hive e HBase.



Introdução ao Apache Spark

A verdade é que criaram o Hadoop para processar grandes volumes de dados em batch.

Mas e se o volume de dados não for tão grande assim?

E se o volume de dados estiver em streaming, ou seja, fluxo contínuo de dados?

O Hadoop MapReduce possui limitações e não atende a alguns requisitos cada vez mais importantes:

- Programação iterativa (Machine Learning, Algoritmos, etc...)
- E streaming de dados

Introdução ao Apache Spark

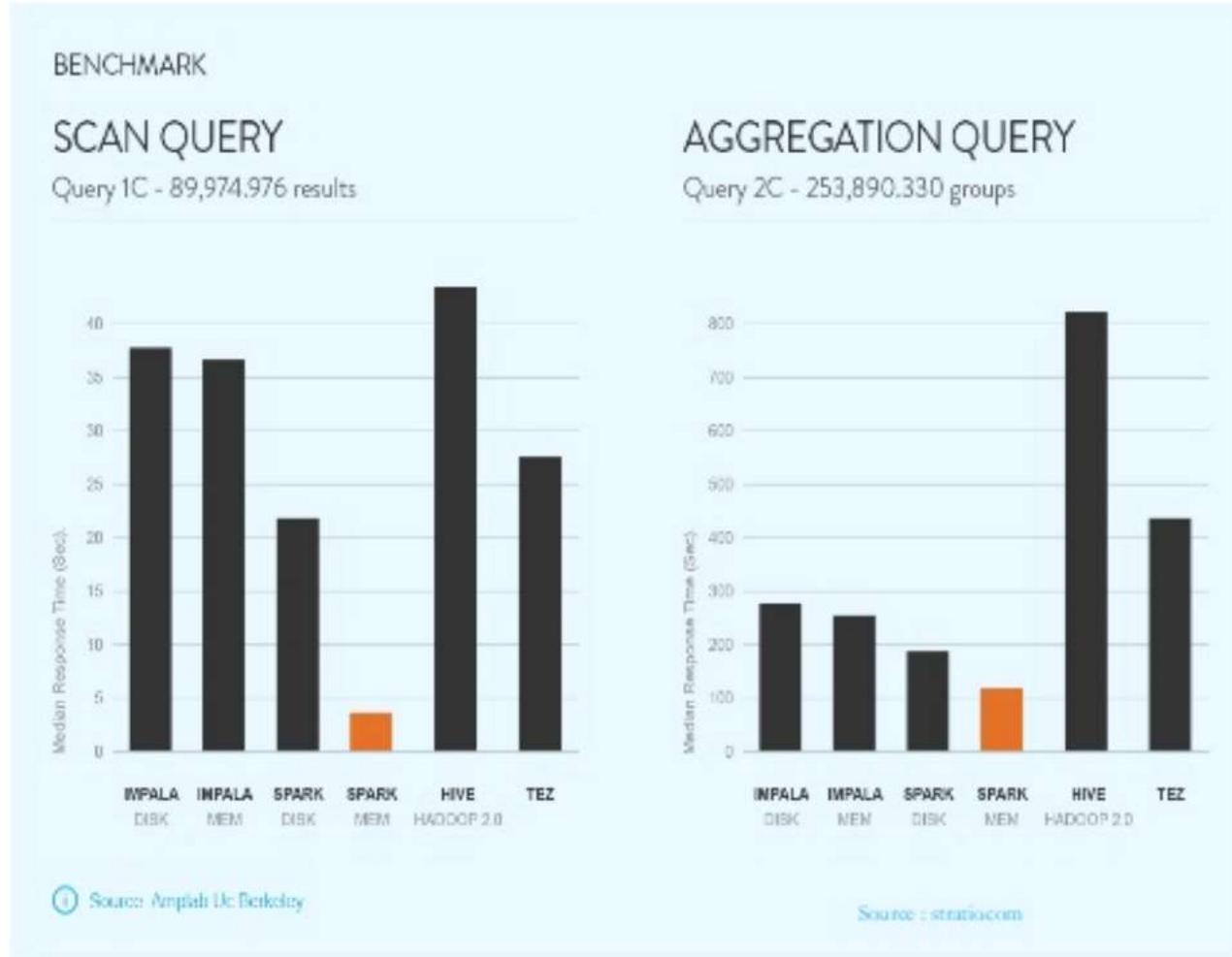


Engine de computação em cluster

- **Veloz** – em memória os dados são processados ate 100x mais rápido que no MapReduce
- **Propósito geral** – SQL, Streaming, Machine Learning
- **Compatibilidade** – Hadoop, Mesos, Yarn, Standalone, HDFS, S3, Cassandra, HBase
- **Mais fácil e simples**

É a primeira plataforma de Big Data a integrar batch, streaming e computação interativa em um único framework

Introdução ao Apache Spark



Introdução ao Apache Spark

| Hadoop | Spark |
|--|---|
| Armazenamento distribuído + Computação distribuída | Somente computação distribuída |
| Framework MapReduce | Computação genérica |
| Normalmente processa dados em disco (HDFS) | Em disco / Em memória |
| Não é ideal para trabalho iterativo | Excelente para trabalhos iterativos (Machine Learning) |
| Processo batch | Até 10x mais rápido para dados em disco Até 100x mais rápido para dados em memória |
| Basicamente Java | Suporta Java, Python, Scala |
| Não possui um shell unificado | Shell para exploração ad-hoc |

Introdução ao Apache Spark



Então o Spark vai substituir o Hadoop

Não.

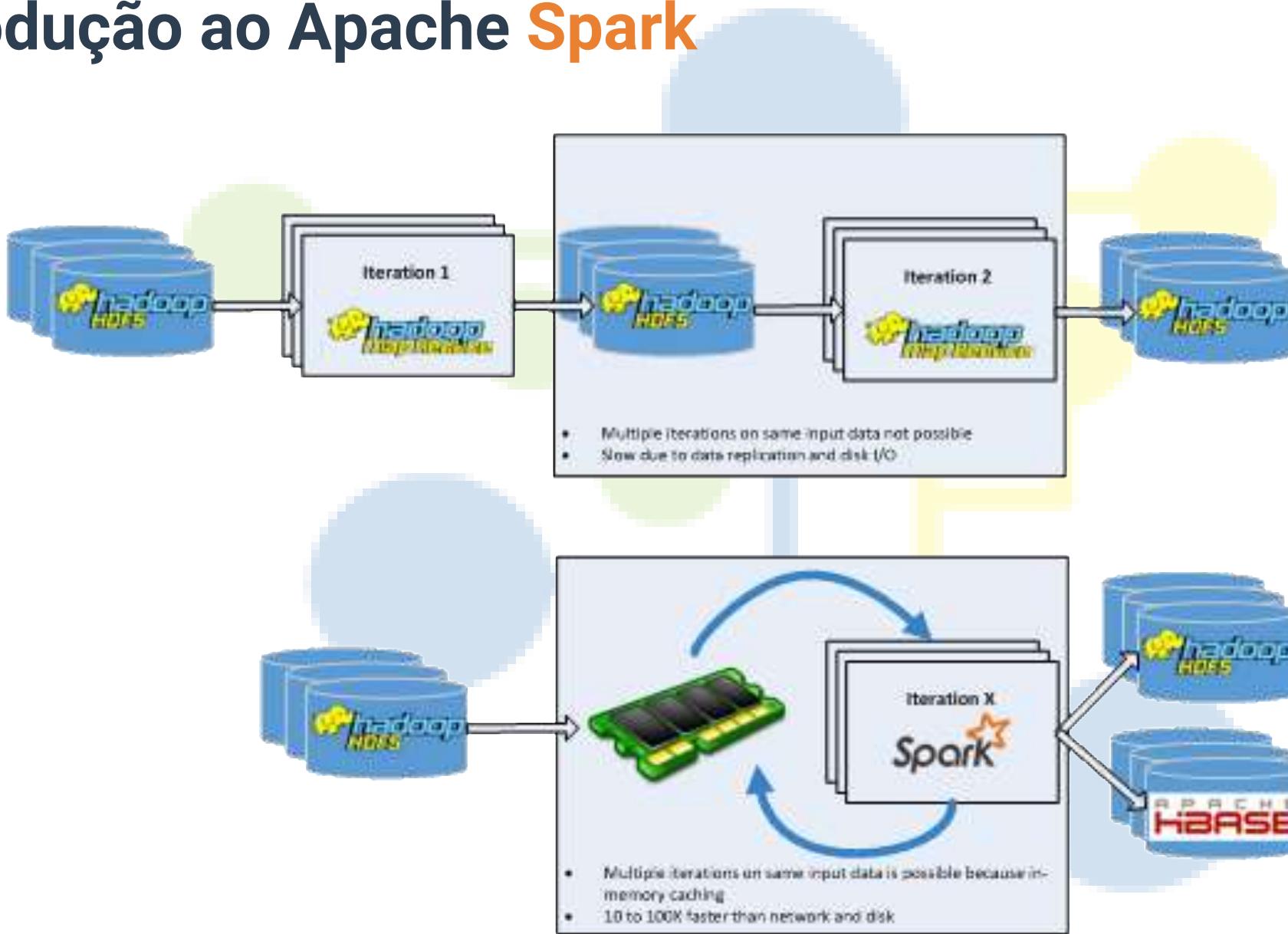
Introdução ao Apache Spark

Então o Spark vai substituir o Hadoop

Não.

- Spark executa sobre o HDFS / YARN
- Pode acessar o HDFS
- Usa YARN para gerenciamento do cluster
- Spark é realmente bom quando os dados podem ser processados em memória
Mas e quando não podem (por exemplo, gigantescos volumes de dados)?

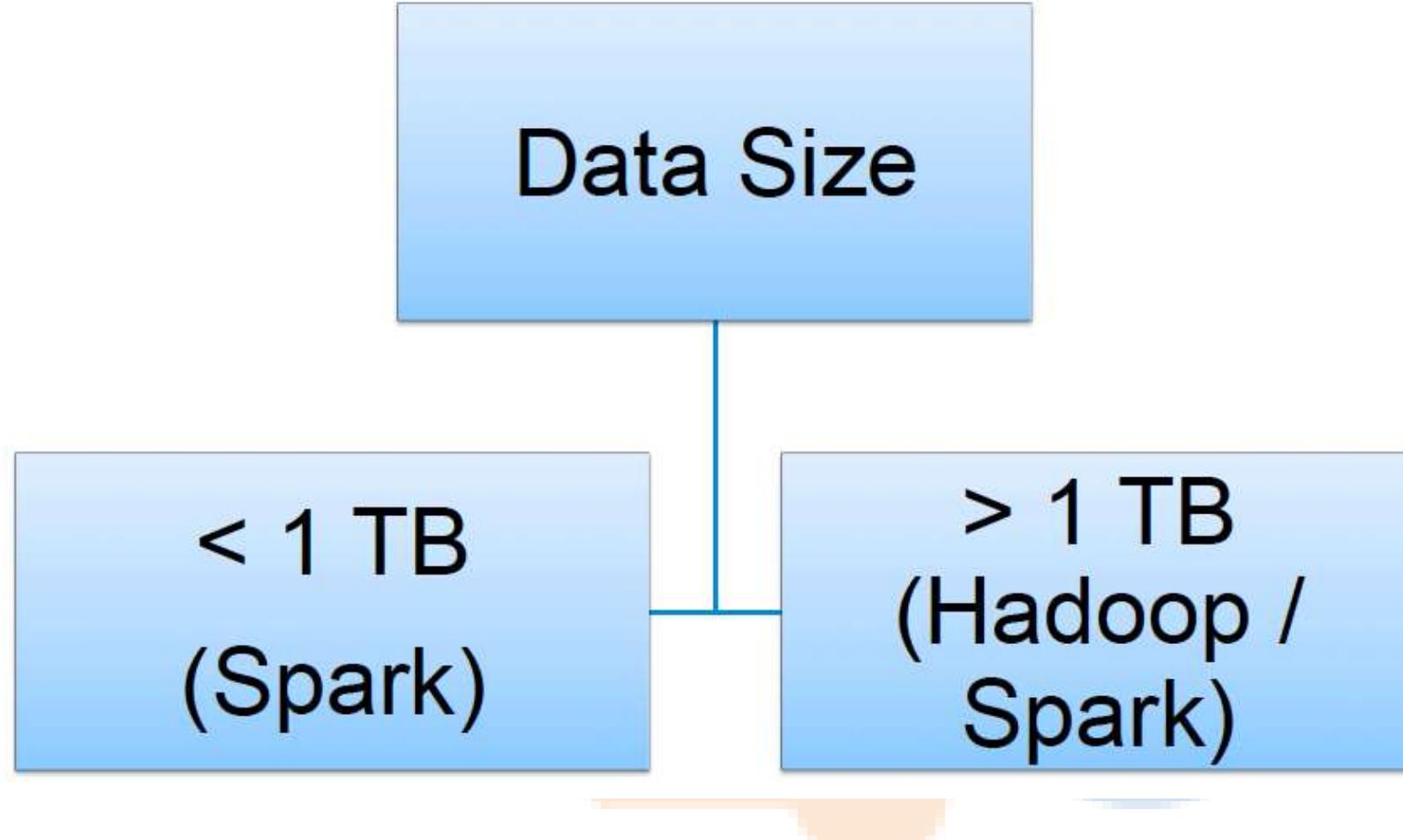
Introdução ao Apache Spark



Introdução ao Apache Spark

| | Hadoop | Spark |
|--|------------------------------------|---|
| Processamento batch | Hadoop MapReduce (Java, Pig, Hive) | Spark RDD (Java, Python, Scala) |
| Query SQL | Hadoop: Hive | Spark SQL |
| Processamento Stream / Processamento em Tempo Real | Storm, Kafka | Spark Streaming |
| Machine Learning | Mahout | Spark ML Lib |
| Algoritmos iterativos | Lento | Muito rápido (em memória) |
| Workflow ETL | Pig, Flume | Pig com Spark ou Mix de Spark SQL e programação RDD |
| Volume de Dados | Volume gigante (Petabytes) | Volume médio (Gigabytes / Terabytes) |

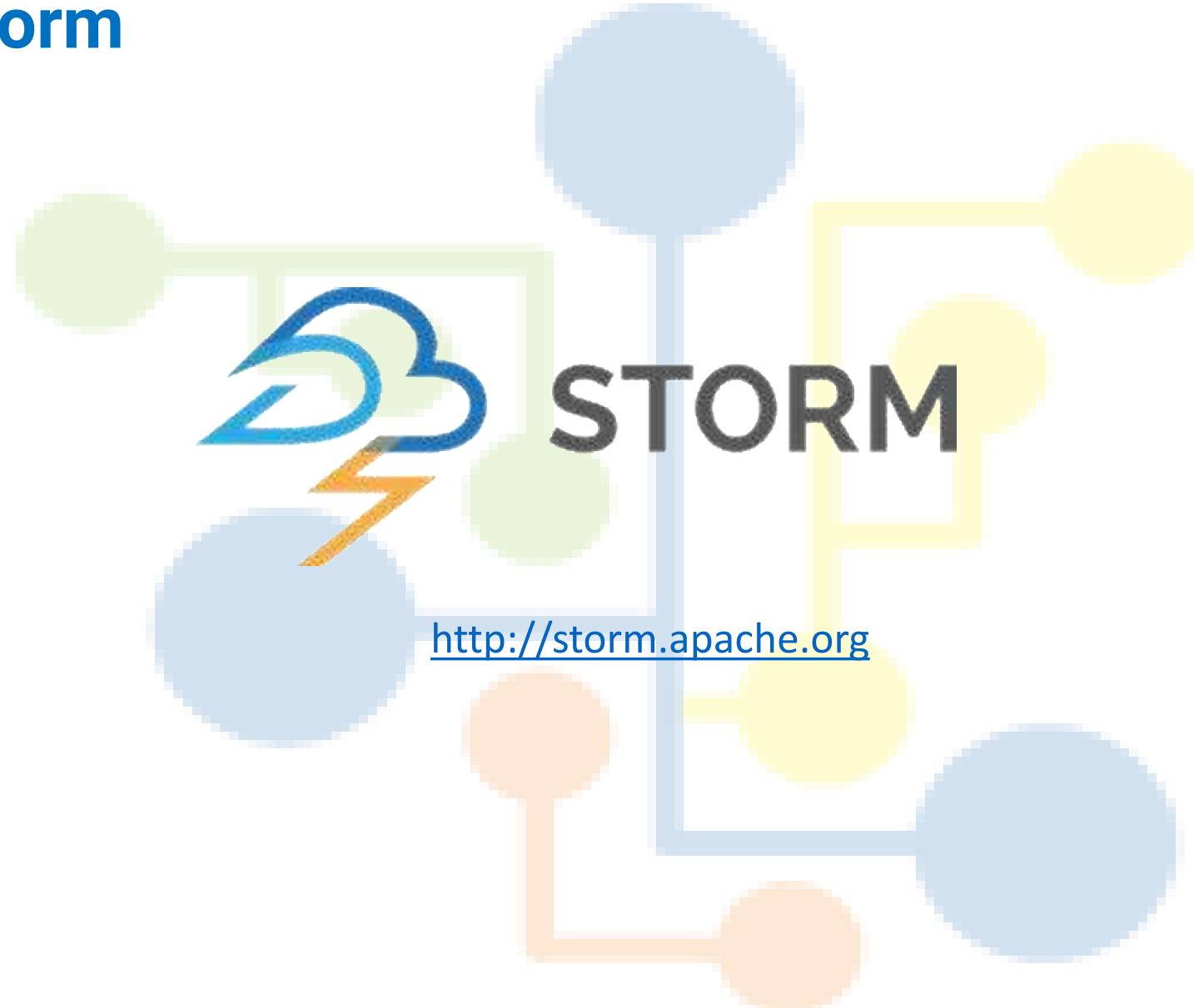
Introdução ao Apache **Spark**



Introdução ao Apache Spark



Apache Storm



<http://storm.apache.org>

Apache Storm



O Apache Storm se tornou o padrão para processamento em tempo real distribuído e permite processar grandes quantidades de dados

Apache Storm



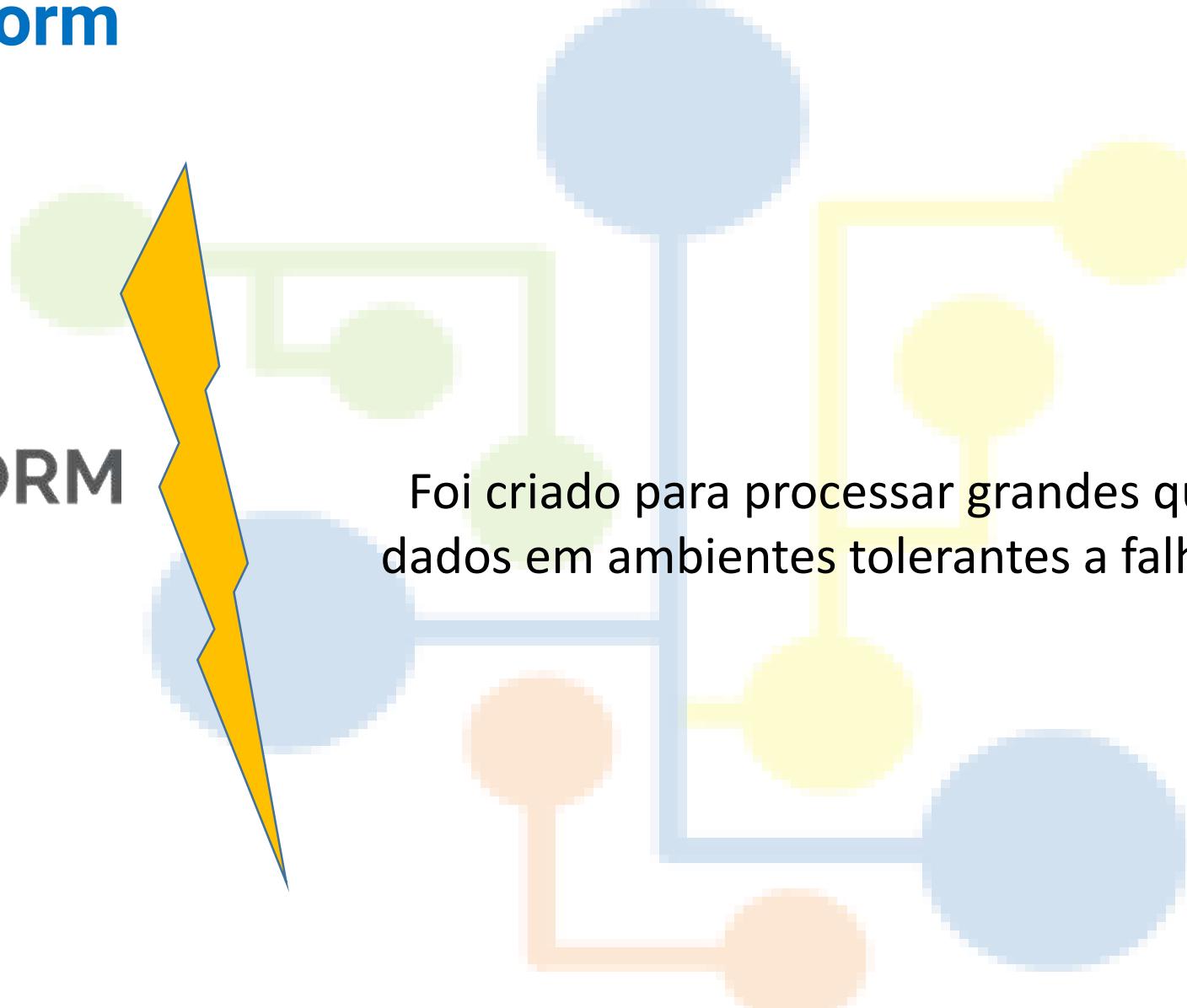
STORM

O Apache Storm foi desenvolvido em Java

Apache Storm



STORM



Foi criado para processar grandes quantidades de dados em ambientes tolerantes a falhas e escaláveis

Apache Storm



Basicamente, o Storm é um framework para Streaming de dados (fluxo contínuo de dados) e possui uma alta taxa de ingestão de dados

Apache Storm



STORM

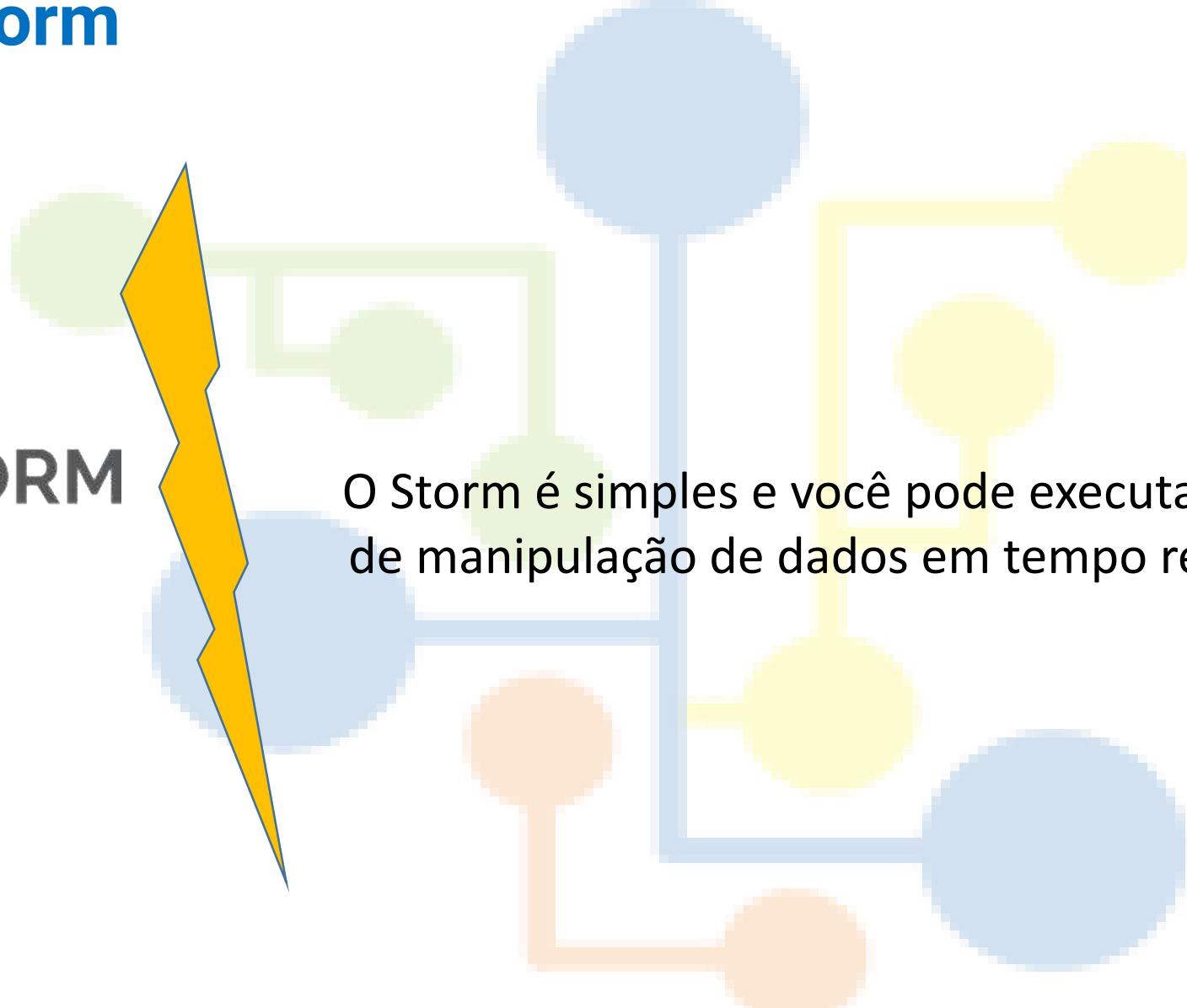


A gestão do estado do cluster é feita através do Zookeeper

Apache Storm



STORM



O Storm é simples e você pode executar todos os tipos de manipulação de dados em tempo real, em paralelo

Apache Storm



STORM

O Apache Storm é um dos líderes em Real-Time Analytics

Principais benefícios de se utilizar o Storm:

- Storm é open-source, robusto e amigável (fácil utilização)
- Tolerante a falhas, flexível, confiável e suporta diversas linguagens de programação
- Processa dados em tempo-real
- Storm é incrivelmente veloz



NOT ONLY
SQL

de Dados

APACHE
HBASE

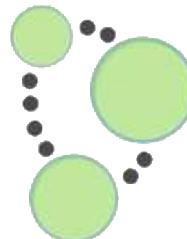


CouchDB
relax

riak

The logo for riak consists of a stylized network icon followed by the word "riak" in a large, lowercase, sans-serif font.

HYPERTABLE INC



Neo4j



redis

Cassandra

The logo for Cassandra features a blue eye with a sunburst iris.

mongoDB

The logo for mongoDB consists of a green leaf icon followed by the word "mongoDB" in a lowercase, sans-serif font.

Banco de Dados NoSQL

Bancos de Dados tradicionais RDBMS (Relational Database Management Systems) são projetados para tratar grandes quantidades de dados não-estruturados (Big Data)



Banco de Dados NoSQL

Bancos de Dados tradicionais foram projetados somente para tratar conjuntos de dados que possam ser armazenados em linhas e colunas e portanto, possam ser consultados através do uso de queries utilizando linguagem SQL (Structured Query Language)



Banco de Dados NoSQL

Bancos de Dados relacionais não
são capazes de tratar dados não-
estruturados ou semi-estruturados



Banco de Dados NoSQL

Ou seja, Bancos de Dados relacionais simplesmente não possuem funcionalidades necessárias para atender os requisitos do Big Data, dados gerados em grande volume, alta velocidade e variedade



Banco de Dados NoSQL

Esta lacuna está sendo preenchida por
Bancos de Dados NoSQL



Banco de Dados NoSQL



Bancos de Dados NoSQL, são bancos de dados distribuídos e não-relacionais, que foram projetados para atender os requerimentos do Big Data

Banco de Dados NoSQL



Bancos de Dados NoSQL oferecem uma arquitetura muito mais escalável e eficiente que os bancos relacionais e facilitam consultas no-sql de dados semi-estruturados ou não-estruturados

Banco de Dados NoSQL

Existe alguma discussão sobre o significado de NoSQL.

Alguns afirmam que a sigla significa *Not Only SQL*, enquanto outros afirmam que significa *Non-SQL*. Não há um consenso sobre isso. Mas pense sobre NoSQL como uma classe de banco de dados não-relacionais que não se enquadram na classificação de bancos de dados relacionais (RDBMS), que utilizam linguagem SQL.



Banco de Dados NoSQL



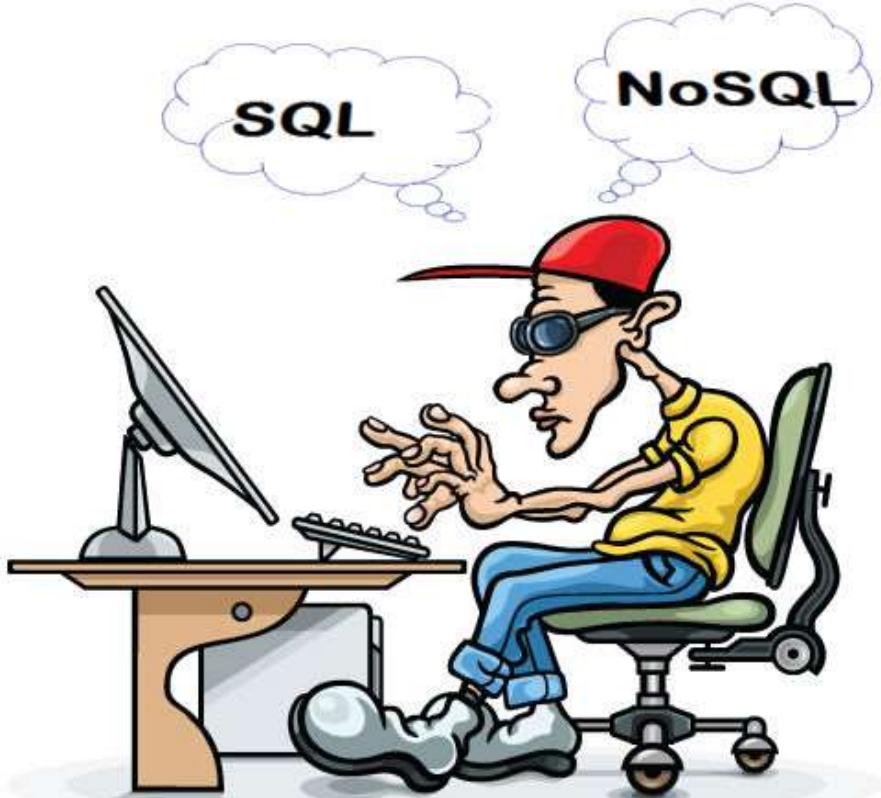
Embora o modelo relacional e a Structured Query Language (SQL) foram por décadas o padrão para armazenamento de dados, é fato que os bancos de dados relacionais não são mais os vencedores quando se trata de flexibilidade e escalabilidade

Banco de Dados NoSQL



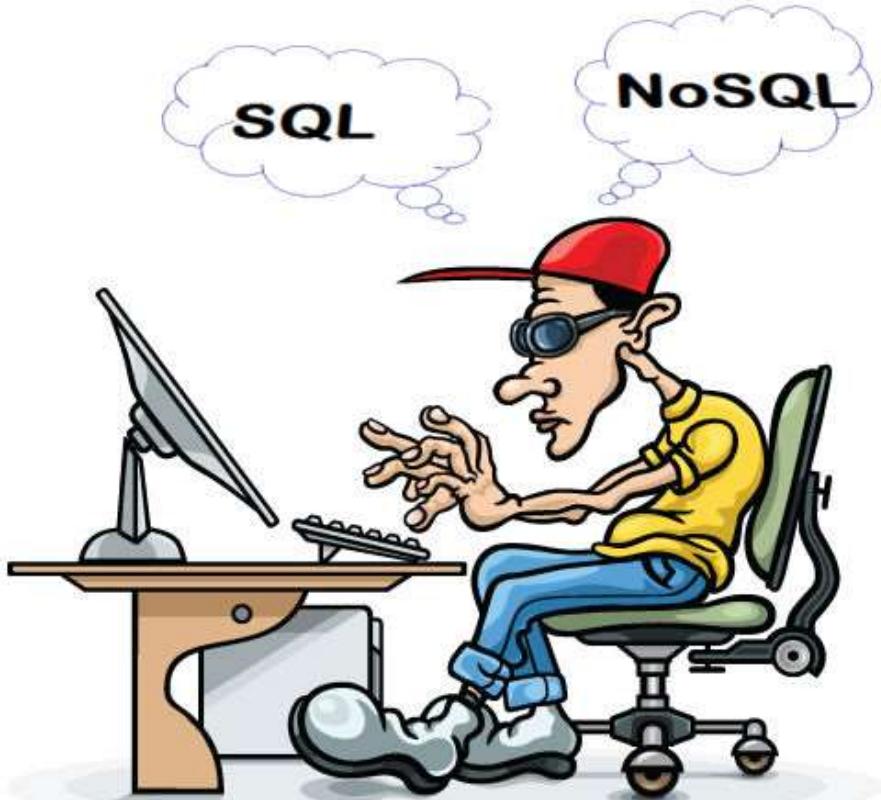
Isto tornou-se verdadeiro especialmente com o advento das redes sociais e Internet das Coisas

Banco de Dados NoSQL



A este respeito, NoSQL surgiu como um paradigma não-tradicional para lidar com grandes volumes de dados e para resolver os desafios colocados pela chegada de implementações de Big Data

Banco de Dados NoSQL



Atualmente, bancos de dados NoSQL como MongoDB, Cassandra e CouchDB introduzem novas características e funcionalidades, trazendo ainda mais inovação e resultados surpreendentes

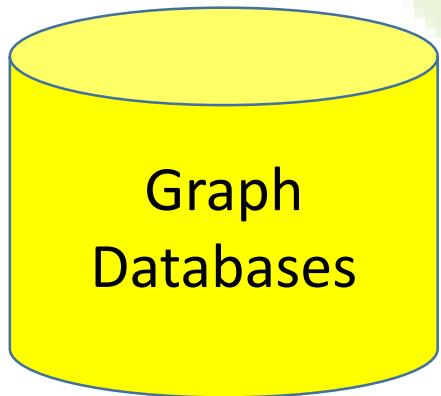
Banco de Dados NoSQL



Bancos de Dados NoSQL oferecem 4 categorias principais de bancos de dados:

- Graph databases
- Document databases
- Key-values stores
- Column family stores

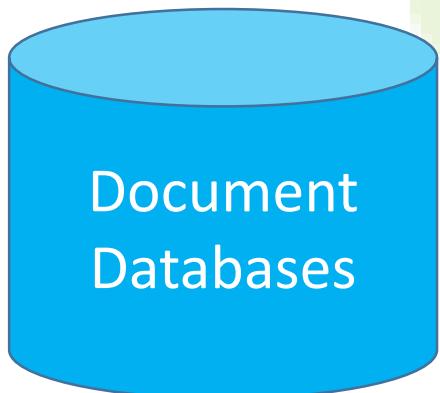
Banco de Dados NoSQL



Esta categoria de Bancos de Dados NoSQL, geralmente é aderente a cenários de rede social online, onde os nós representam as entidades e os laços representam as interconexões entre eles.

Desta forma, é possível atravessar o grafo seguindo as relações. Esta categoria têm sido usada para lidar com problemas relacionados a sistemas de recomendação e listas de controle de acesso, fazendo uso de sua capacidade de lidar com dados altamente interligados.

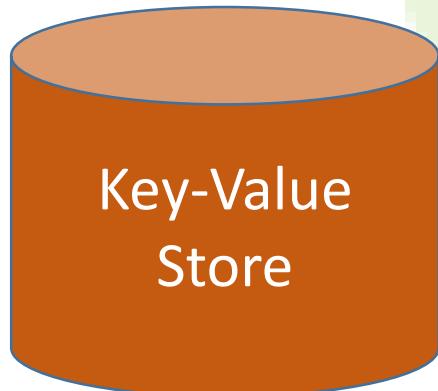
Banco de Dados NoSQL



Esta categoria de Bancos de Dados NoSQL permite o armazenamento de milhões de documentos.

Por exemplo, você pode armazenar detalhes sobre um empregado, junto com o currículo dele (como um documento) e então pesquisar sobre potenciais candidatos a uma vaga, usando um campo específico, como telefone ou conhecimento em uma tecnologia.

Banco de Dados NoSQL



Key-Value
Store

Nesta categoria, os dados são armazenados no formato key-value (chave-valor) e os valores (dados) são identificados pelas chaves.

É possível armazenar bilhões de registros de forma eficiente e o processo de escrita é bem rápido. Os dados podem ser então pesquisados através das chaves associadas.

Banco de Dados NoSQL

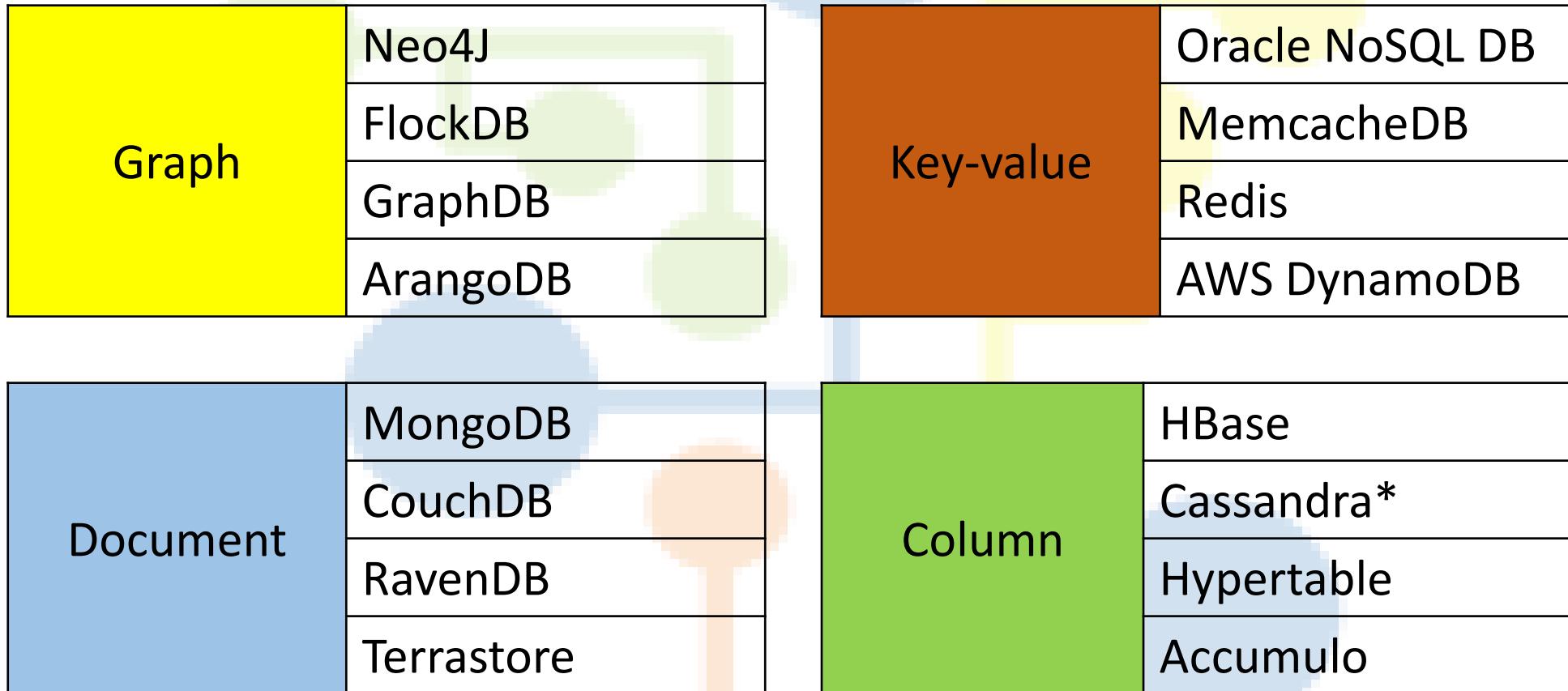


Também chamados bancos de dados orientados a coluna, os dados são organizados em grupos de colunas e tanto o armazenamento, quanto as pesquisas de dados são baseados em chaves.

HBase e Hypertable são os exemplos mais comuns desta categoria.

Banco de Dados NoSQL

Os principais Bancos de Dados NoSQL são:



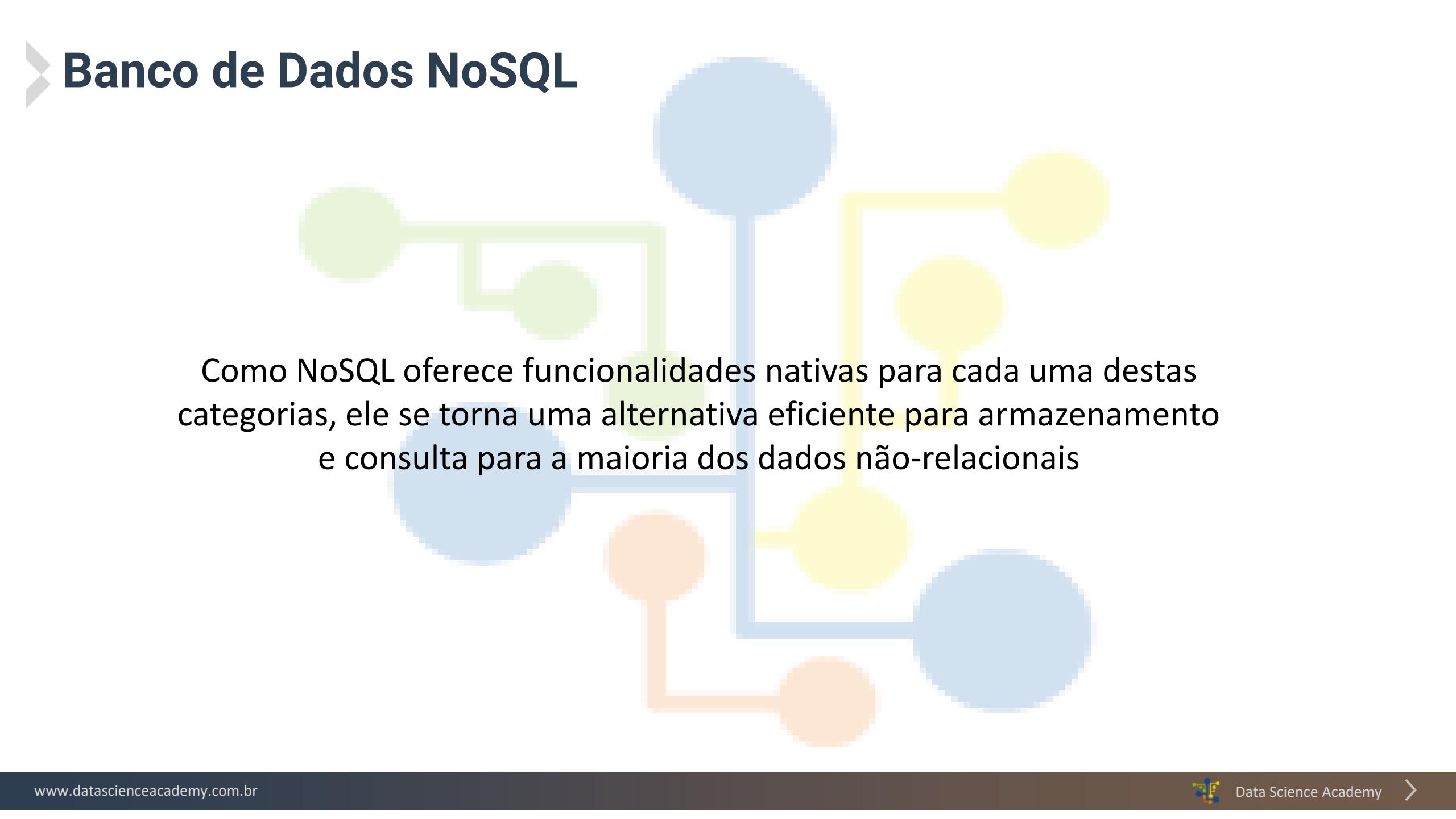
Banco de Dados NoSQL



Para uma lista completa de Bancos de Dados NoSQL visite:

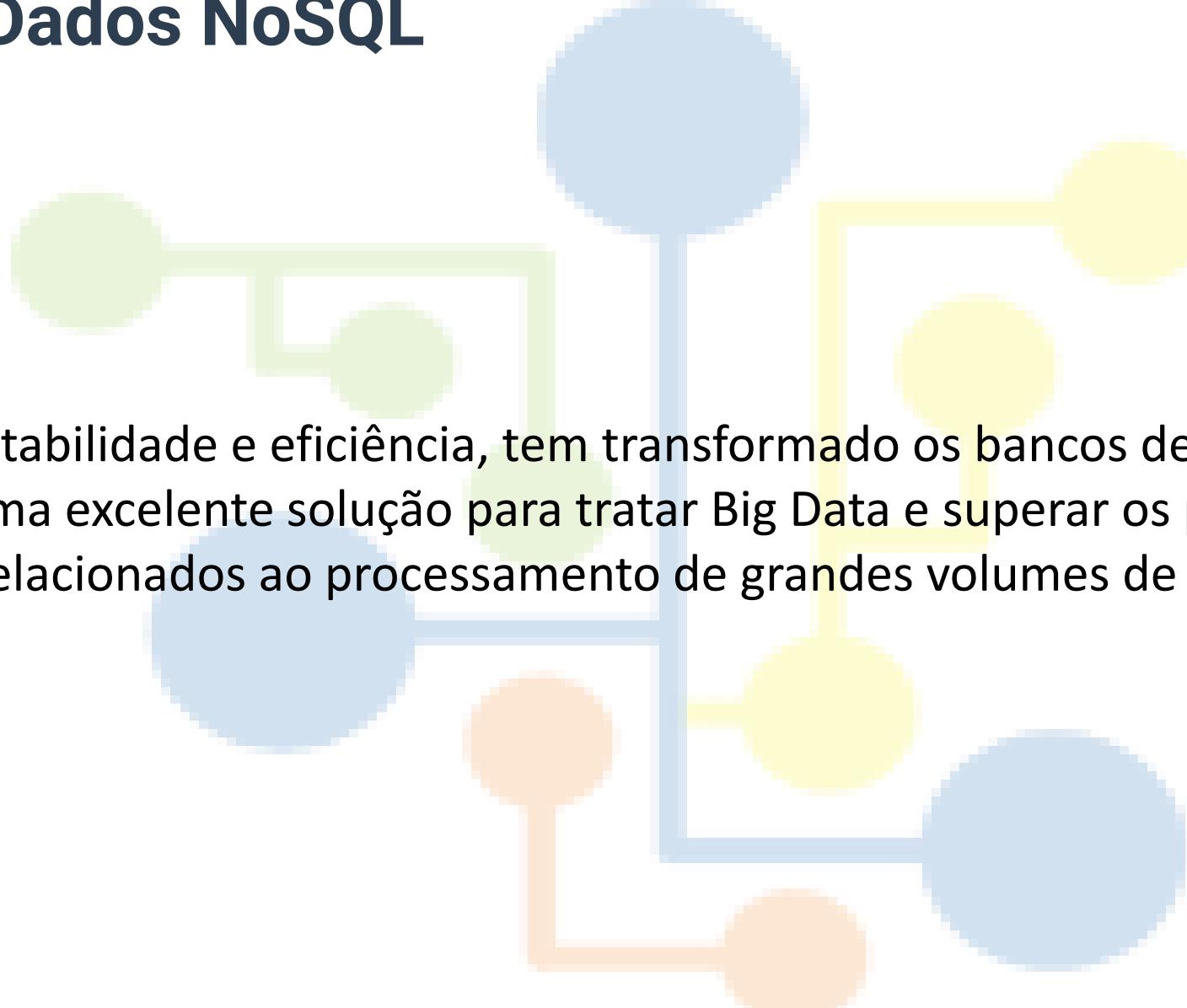
<http://nosql-database.org>

Banco de Dados NoSQL



Como NoSQL oferece funcionalidades nativas para cada uma destas categorias, ele se torna uma alternativa eficiente para armazenamento e consulta para a maioria dos dados não-relacionais

Banco de Dados NoSQL



Esta adaptabilidade e eficiência, tem transformado os bancos de dados NoSQL em uma excelente solução para tratar Big Data e superar os problemas relacionados ao processamento de grandes volumes de dados

Banco de Dados NoSQL

E por que usar bancos de dados NoSQL?

- Representação de dados sem esquema
- Tempo de desenvolvimento
- Velocidade
- Escalabilidade



Banco de Dados NoSQL



MongoDB é um banco de dados orientado a documento, uma das categorias de bancos de dados NoSQL.

Banco de Dados NoSQL



Um banco de dados NoSQL orientado a documento, substitui o conceito de "linha" como em bancos de dados relacionais, por um modelo mais flexível, o "documento".

Banco de Dados NoSQL



o MongoDB é open-source e um dos líderes no segmento de bancos de dados NoSQL. Ele foi desenvolvido em linguagem C++.

Banco de Dados NoSQL



Algumas das principais características do MongoDB:

- Indexação

O MongoDB suporta índices secundários, permitindo a construção de queries mais velozes.

Banco de Dados NoSQL



Algumas das principais características do MongoDB:

- Agregação

O MongoDB permite a construção de agregações complexas de dados, otimizando o desempenho.

Banco de Dados NoSQL



Algumas das principais características do MongoDB:

- Tipos de dados especiais

O MongoDB suporta coleções **time-to-live** para dados que expiram em um determinado tempo, como sessões por exemplo.

Banco de Dados NoSQL



Algumas das principais características do MongoDB:

- Armazenamento

O MongoDB suporta o armazenamento de grandes quantidades de dados.

Banco de Dados NoSQL



Algumas características presentes em bancos de dados relacionais, não estão presentes no MongoDB, como alguns tipos de joins e transações multi-linha.

Banco de Dados NoSQL

| MongoDB | RDBMS |
|--------------------|-------------|
| Database | Database |
| Collection | Table |
| Document | Tuple/Row |
| Field | Column |
| Embedded Documents | Table Join |
| Primary Key | Primary Key |

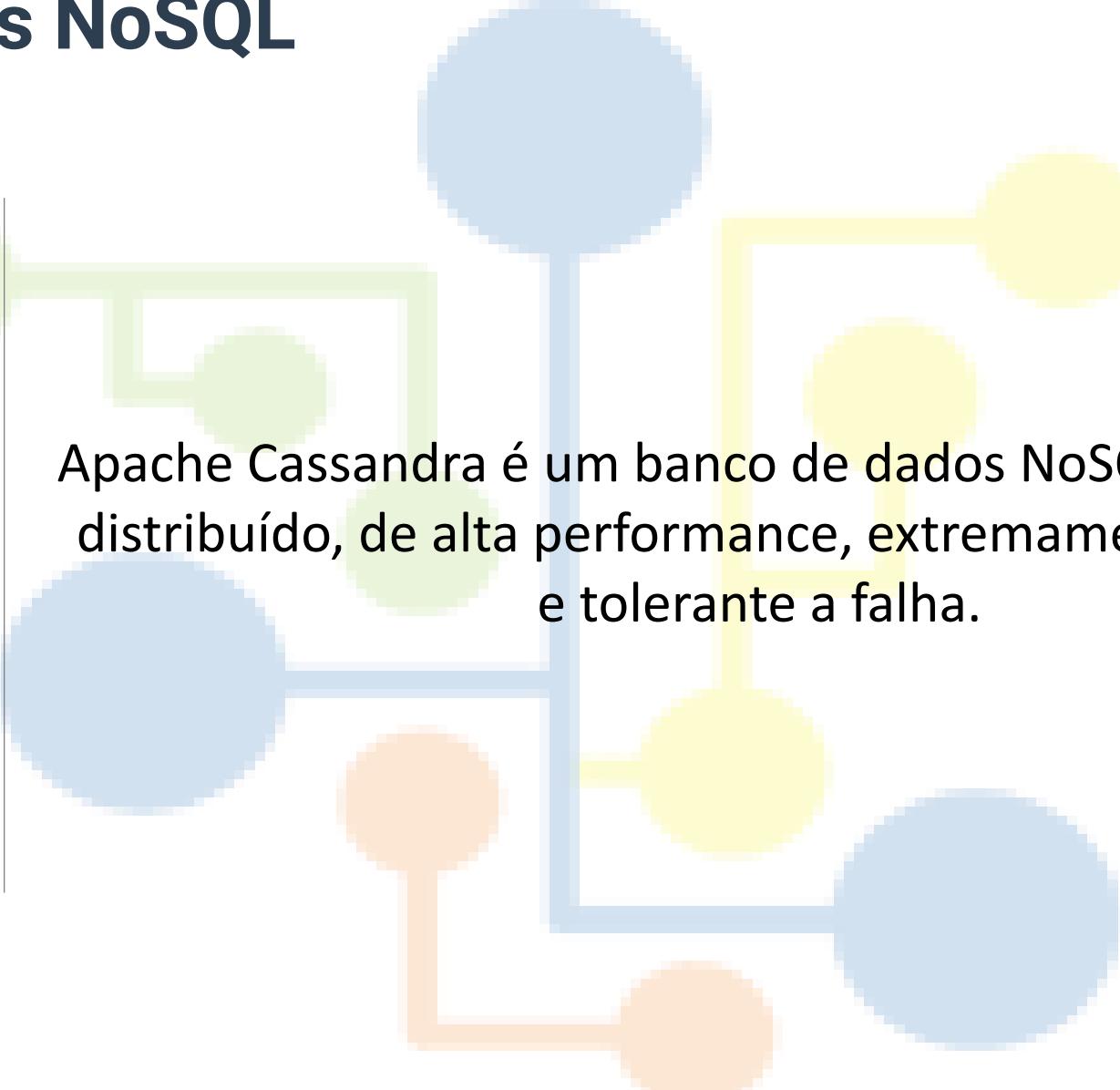
Banco de Dados NoSQL

Onde usar o MongoDB?

- Big Data
- Gestão de Conteúdo
- Infraestrutura Social e Mobile
- Gestão de Dados de Usuários
- Data Hub



Banco de Dados NoSQL



An abstract diagram illustrating a distributed system architecture. It consists of several colored circles (blue, yellow, orange) connected by lines forming a network. Some nodes are enclosed in colored squares (green, blue, yellow), suggesting they represent individual nodes or components within a cluster. The overall shape of the network is roughly rectangular.

Apache Cassandra é um banco de dados NoSQL, livremente distribuído, de alta performance, extremamente escalável e tolerante a falha.

Banco de Dados NoSQL



Ele foi concebido com a premissa que falhas de sistema ou de hardware sempre ocorrem.

Banco de Dados NoSQL



Foi inicialmente desenvolvido pelo Facebook, como uma combinação do BigTable (Google) and Dynamo Data Store (Amazon).

Banco de Dados NoSQL



O Cassandra é usado para armazenar gigantescas quantidades de dados (Big Data), de forma rápida.

Banco de Dados NoSQL



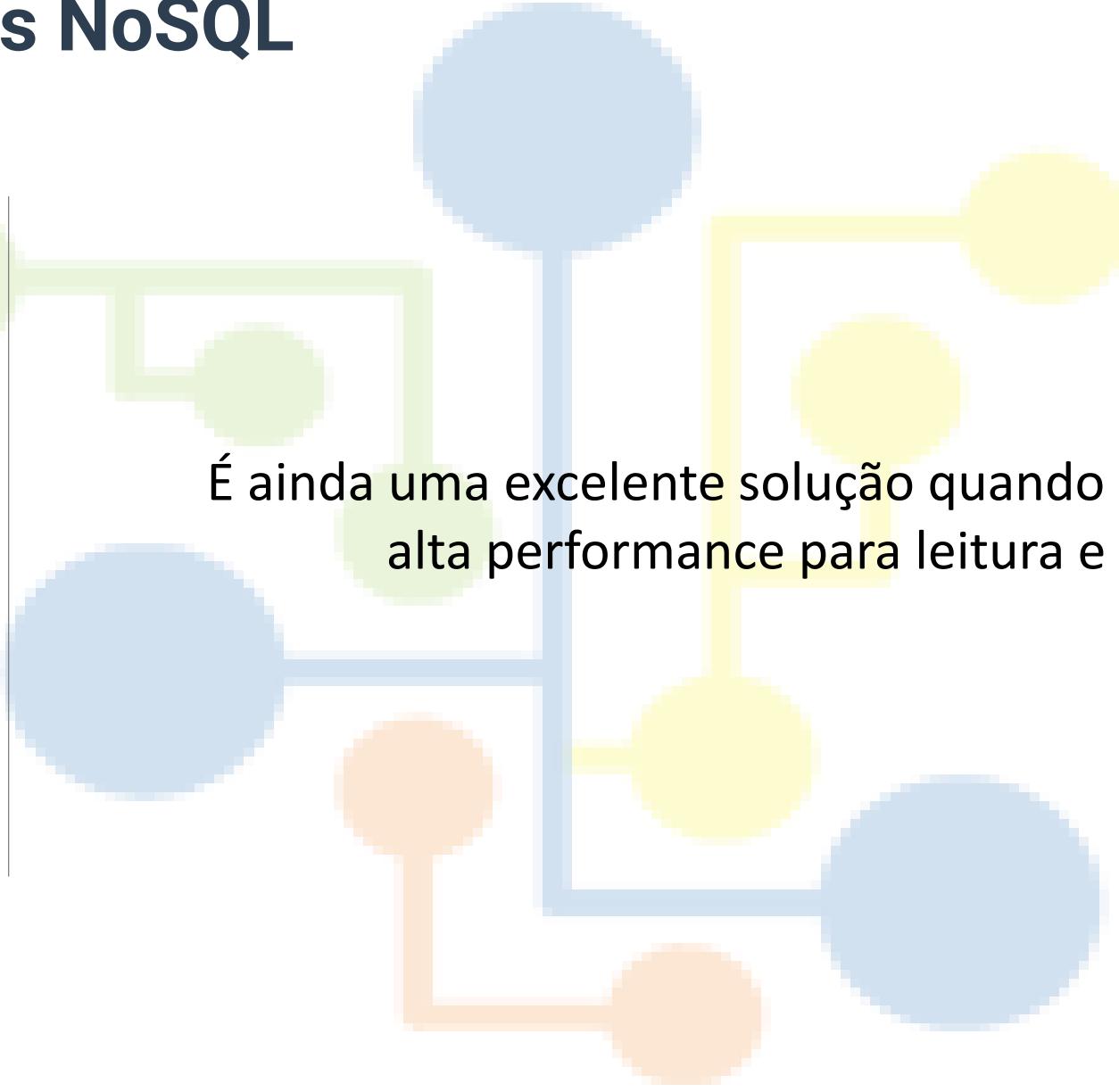
O Cassandra também funciona muito bem quando se faz necessário a pesquisa de dados de forma indexada.

Banco de Dados NoSQL



É voltado para trabalhar em clusters, sendo totalmente escalável. Novos nodes podem ser adicionados, à medida que os dados crescem.

Banco de Dados NoSQL



É ainda uma excelente solução quando se necessita de alta performance para leitura e escrita.

A diagram illustrating a distributed system architecture. It consists of several colored circles (blue, yellow, green, orange) representing nodes, connected by lines forming a network. A central vertical blue line connects four blue nodes at different heights. To the left, a green line connects two green nodes. To the right, a yellow line connects two yellow nodes. At the bottom, an orange line connects two orange nodes. The nodes are semi-transparent, allowing the background to be seen through them.

Banco de Dados NoSQL



Algumas empresas/websites que usam o Cassandra:
eBay, GitHub, GoDaddy, Instagram, Netflix, Reddit,
CERN, Comcast, entre outras.

Banco de Dados NoSQL



Banco de Dados NoSQL



Apache
CouchDB
relax

Banco de Dados NoSQL



Apache
CouchDB
relax



An abstract network diagram composed of various colored circles (blue, green, yellow, orange) connected by lines. It forms a complex web-like structure with some nodes having multiple connections. This visual metaphor represents the distributed nature and interconnectedness of data in a database system.

CouchDB é um banco de dados totalmente voltado para a web.

Banco de Dados NoSQL



No CouchDB os dados são armazenados em documentos JSON (Java Script Object Notation), que consistem em campos que podem ser strings, números, datas, listas ordenadas e mapas associativos.

Banco de Dados NoSQL



Apache
CouchDB
relax

O CouchDB suporta aplicativos web e mobile.

Banco de Dados NoSQL



O CouchDB é distribuído em pares com um server e um client, que podem ter cópias independentes do mesmo banco de dados.

Banco de Dados NoSQL

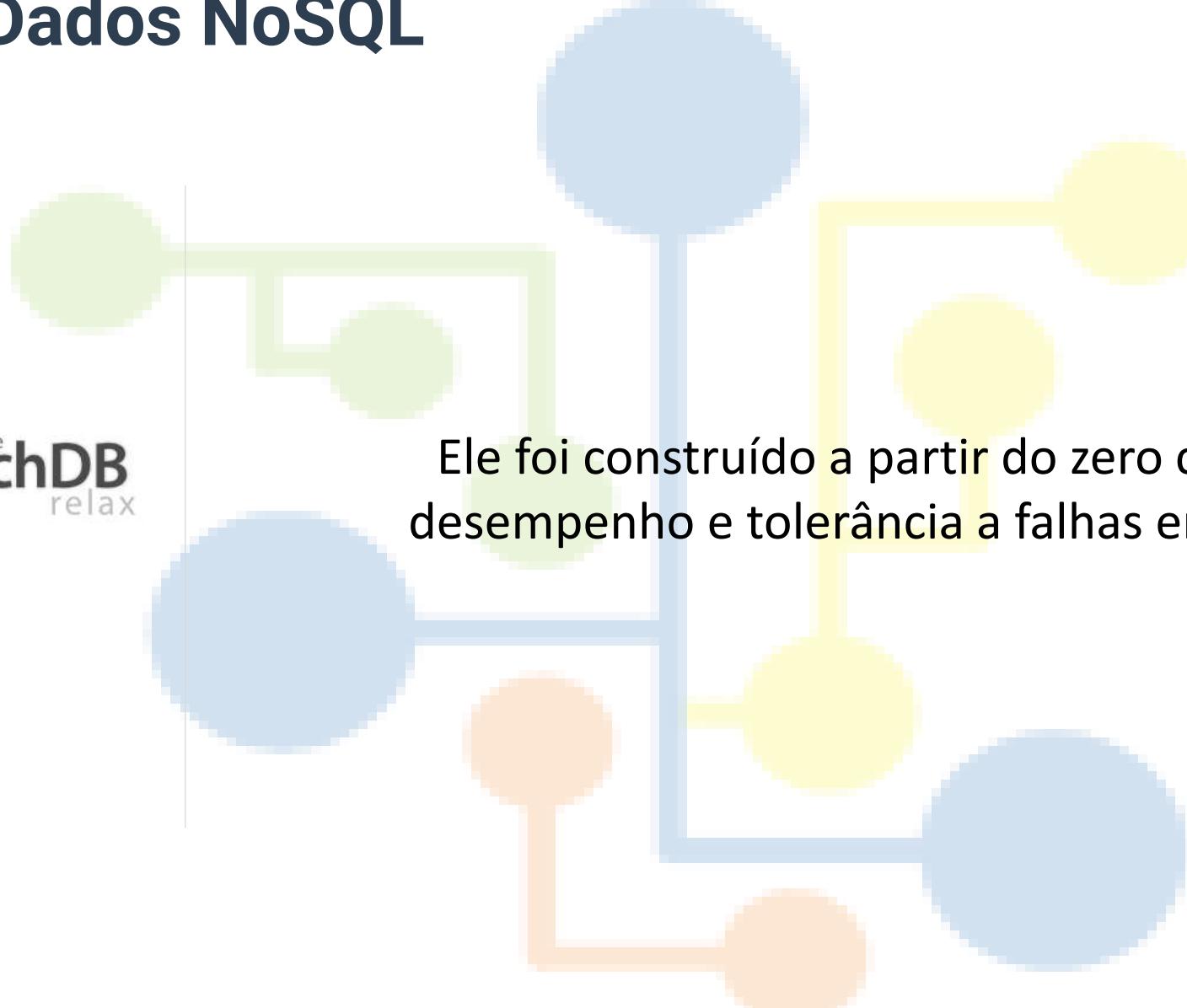


O Apache CouchDB foi o banco de dados que deu o pontapé inicial do movimento NoSQL.

Banco de Dados NoSQL



Apache
CouchDB
relax



Ele foi construído a partir do zero com alto desempenho e tolerância a falhas em mente.

Banco de Dados NoSQL

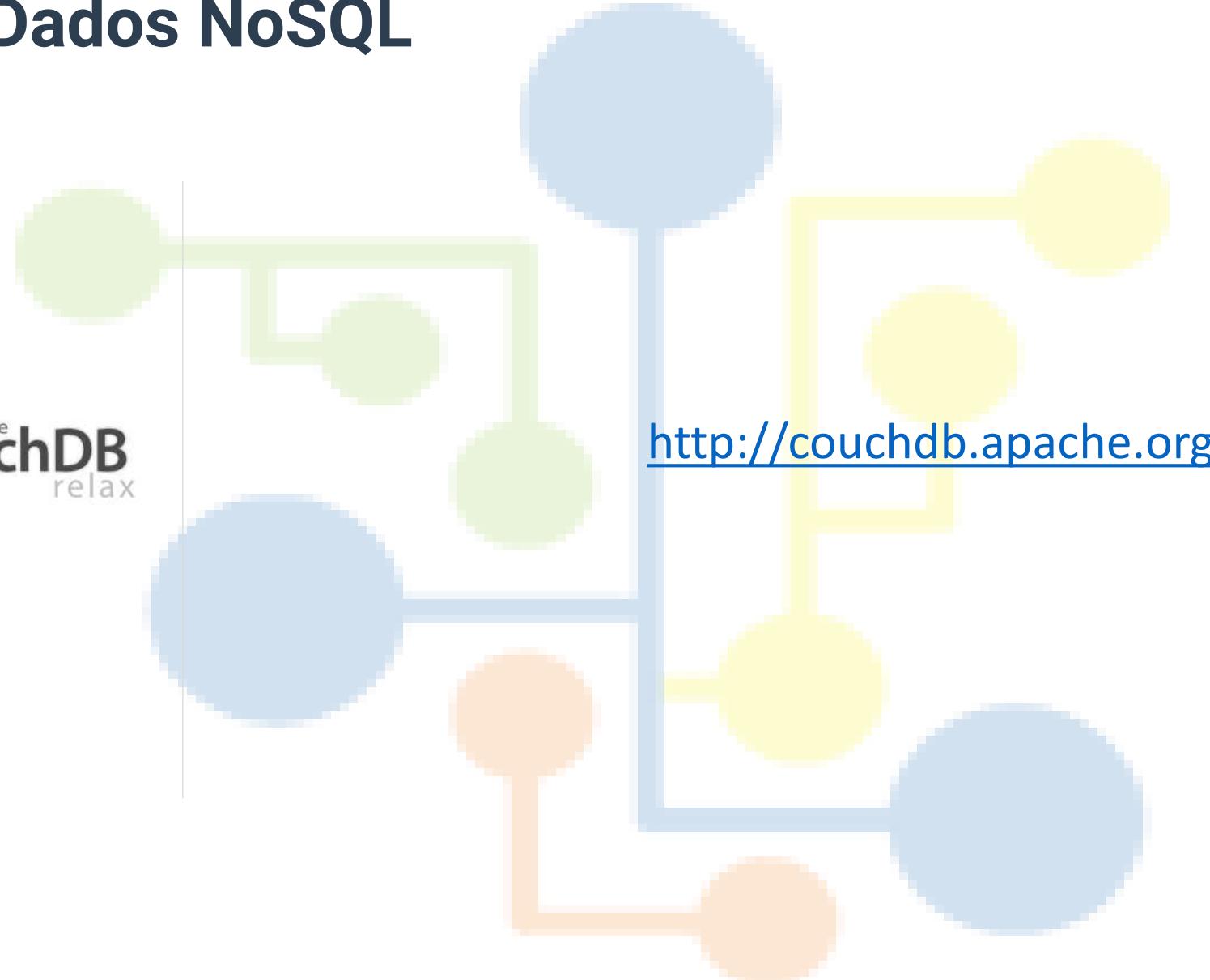


CouchDB permite aos usuários armazenar, reproduzir, sincronizar e processar grandes quantidades de dados (Big Data), distribuídos em dispositivos móveis, servidores, Data Centers e regiões geográficas distintas em qualquer configuração de implantação, incluindo ambiente em nuvem (Cloud).

Banco de Dados NoSQL



Apache
CouchDB
relax



Como as Empresas Estão Utilizando o Big Data?

GLOBAL

DATA CONCEPTS
FUTURE DATA
SOLUTION BUSINESS

The logo consists of the words "SOCIAL", "TEAMWORK", and "IDEA" stacked vertically in a bold, sans-serif font. The letters are white with black outlines, set against a background of a city skyline at night.

INTEGRATED FUTURE
ANALYSIS
INTERNATIONAL
SHARES
PRODUCTIVITY
IDEA
CONF

-BUSINESS

COMMUNICATION VISION

SALES DATA

DATA FINANCE
SHARES TEAM
BUSINESS

Como as Empresas Estão Utilizando o Big Data



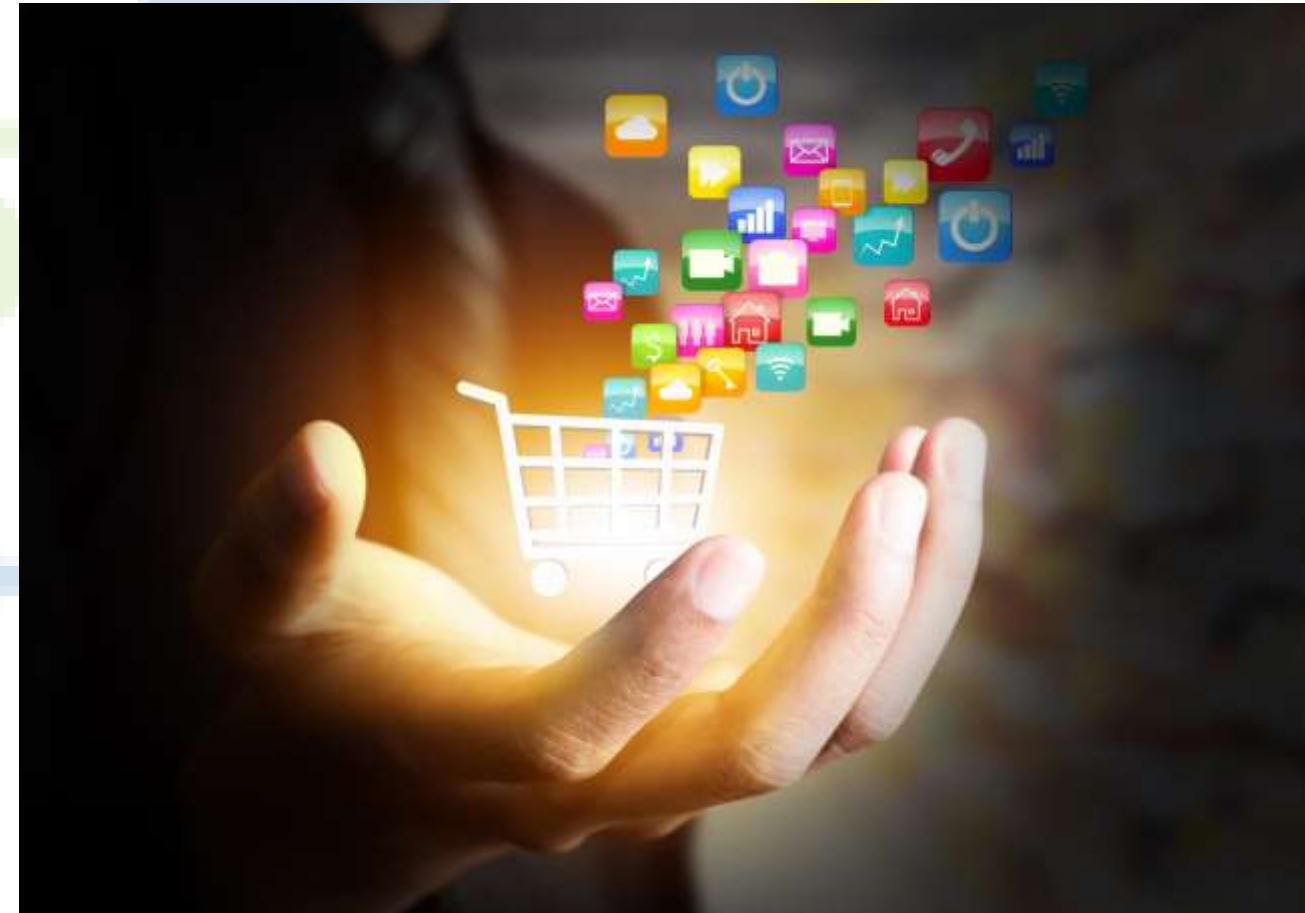
Como as Empresas Estão Utilizando o Big Data



Como as Empresas Estão Utilizando o Big Data



Como as Empresas Estão Utilizando o Big Data



Estão Utile

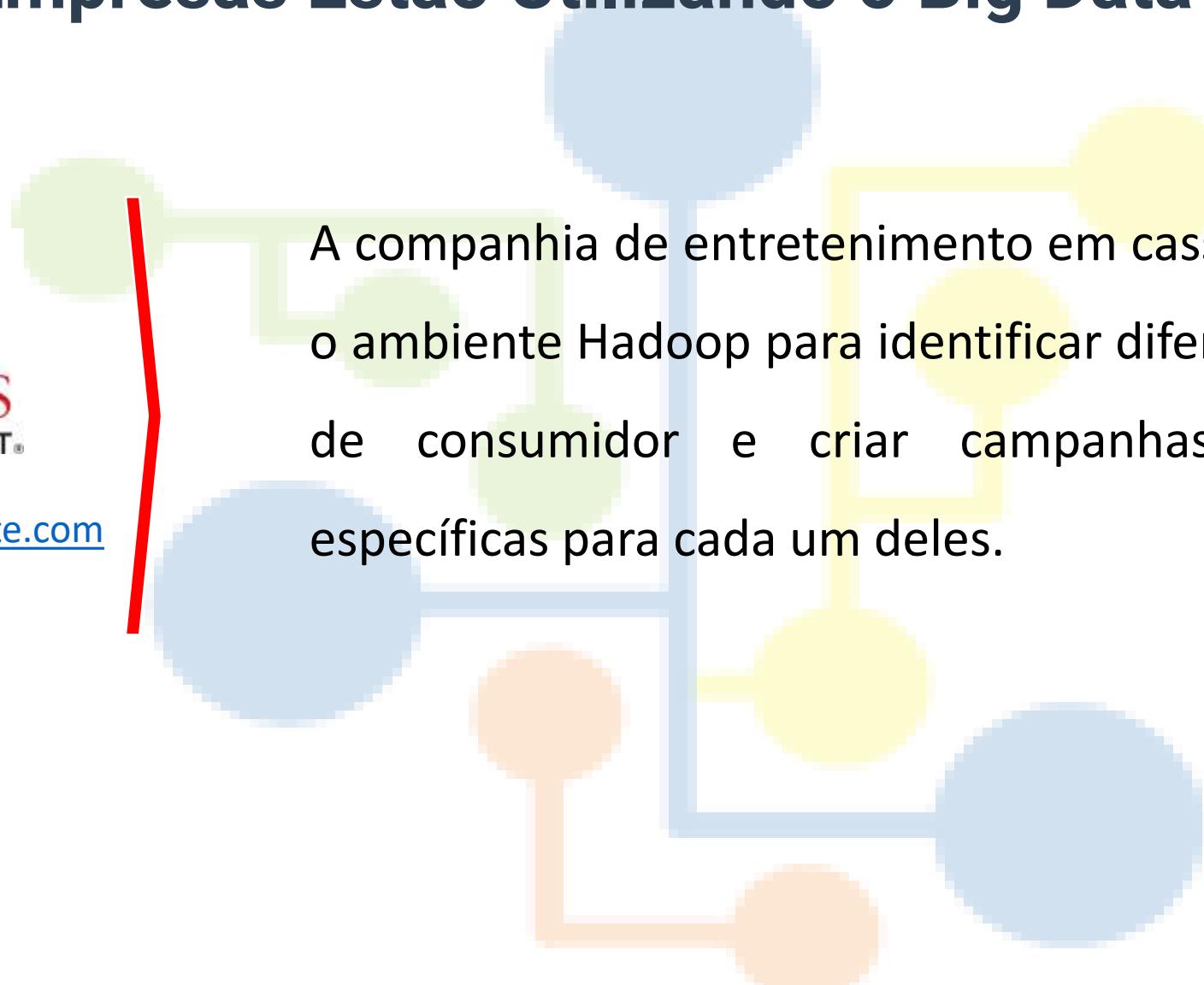


CAESARS
ENTERTAINMENT®

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

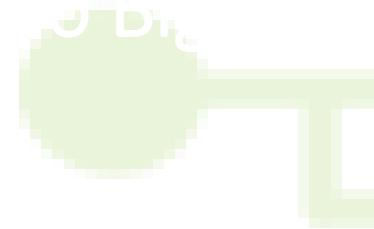


<http://caesarscorporate.com>

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



<http://www.cerner.com>

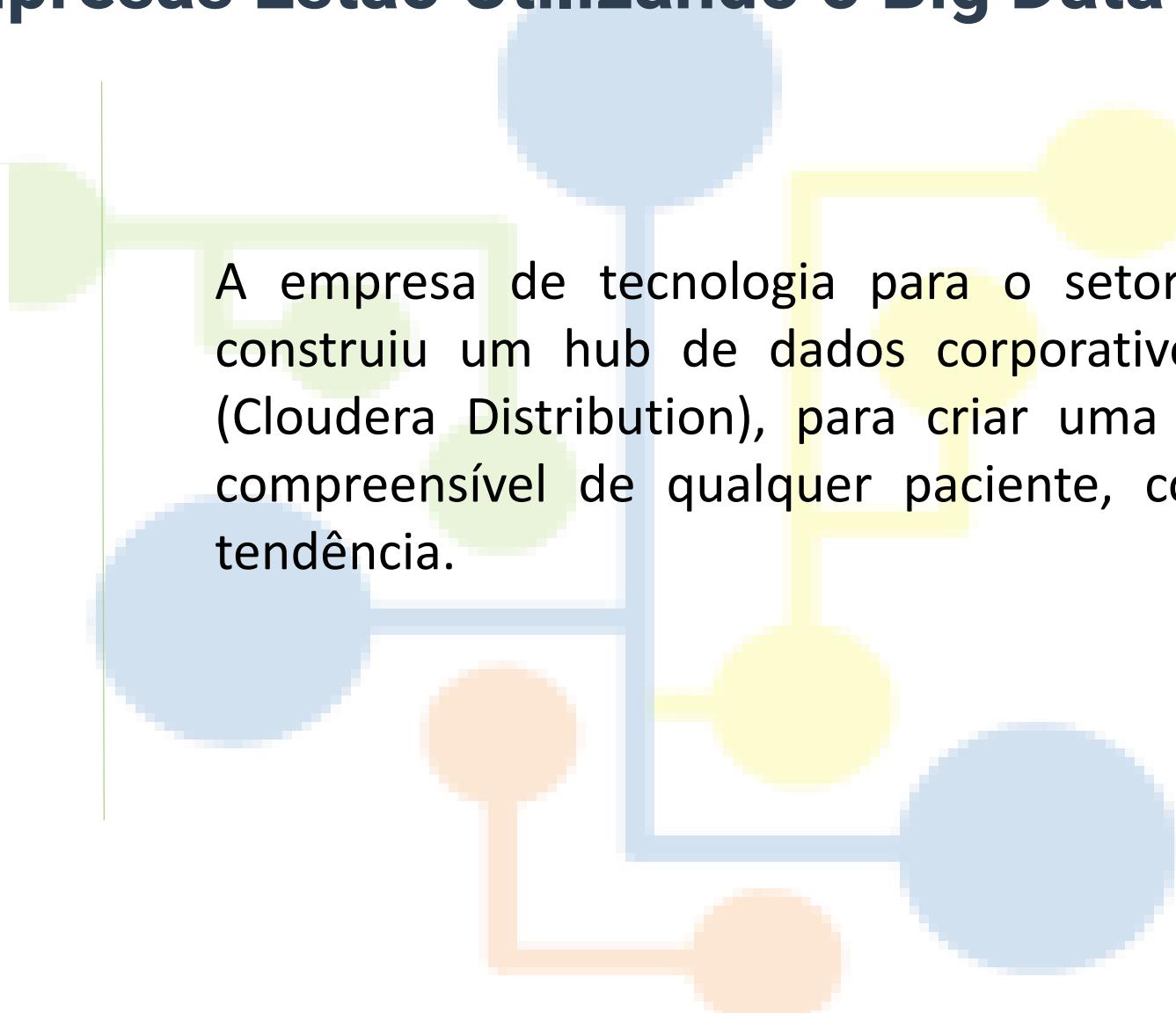


Cerner™

Como as Empresas Estão Utilizando o Big Data



<http://www.cerner.com>

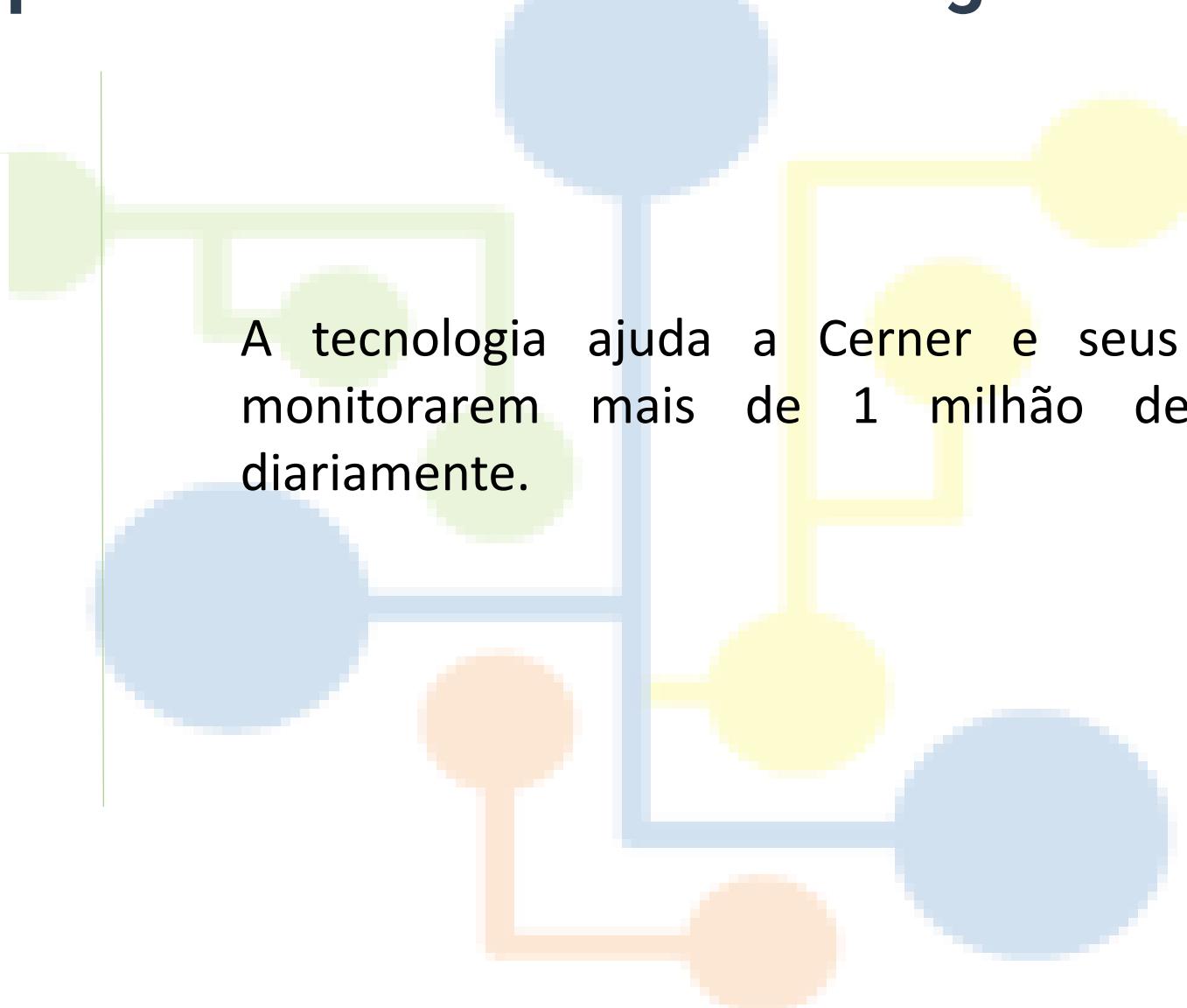
An abstract graphic consisting of several overlapping circles in various colors (blue, yellow, orange) connected by thin lines, forming a network-like structure.

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

Como as Empresas Estão Utilizando o Big Data



<http://www.cerner.com>

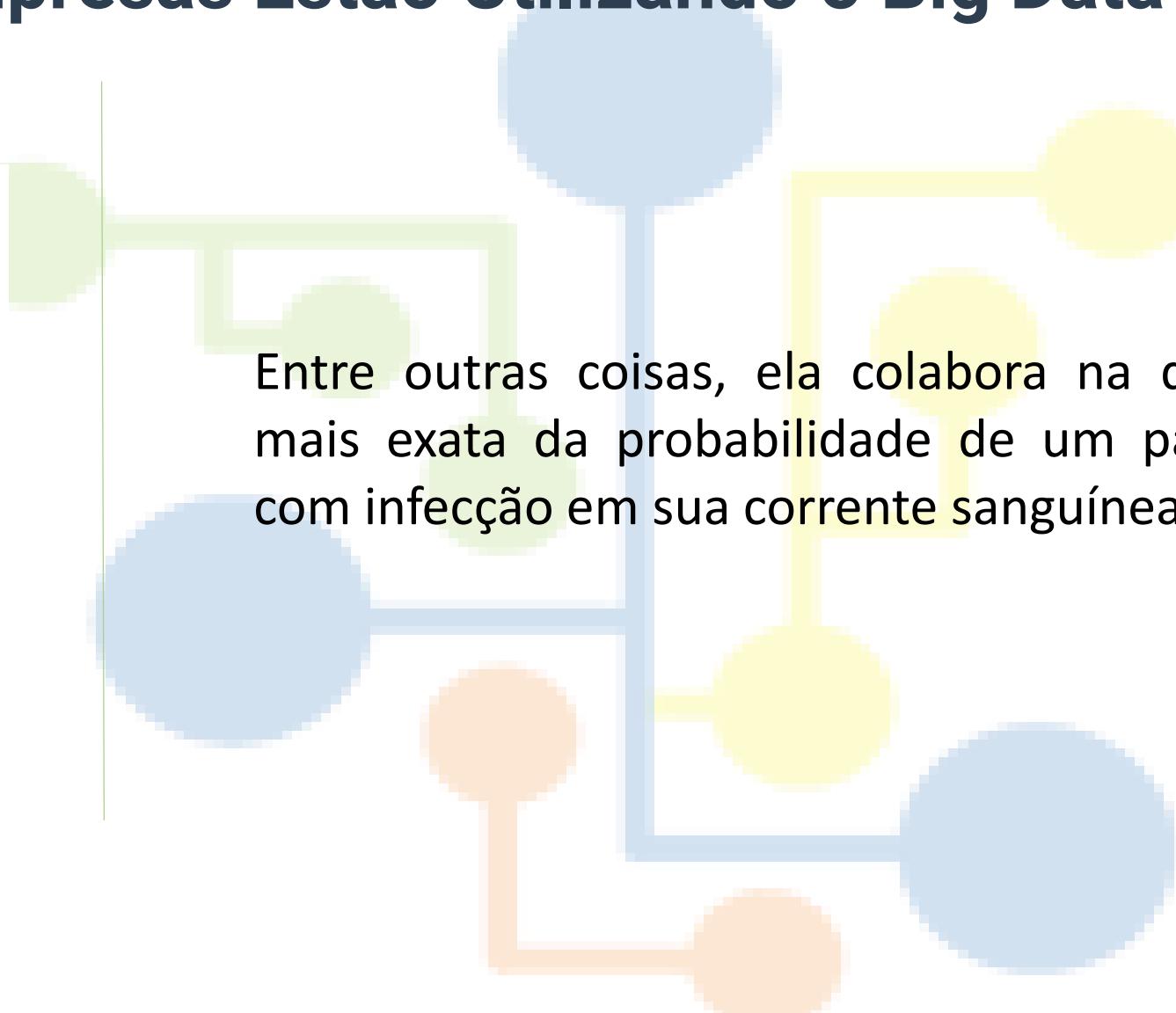


A tecnologia ajuda a Cerner e seus clientes a monitorarem mais de 1 milhão de pacientes diariamente.

Como as Empresas Estão Utilizando o Big Data



<http://www.cerner.com>

An abstract graphic on the right side of the slide features a network of nodes and connecting lines. The nodes are colored circles of various sizes: blue, yellow, orange, and green. They are interconnected by a grid-like structure of thin lines, suggesting a complex data network or system architecture.

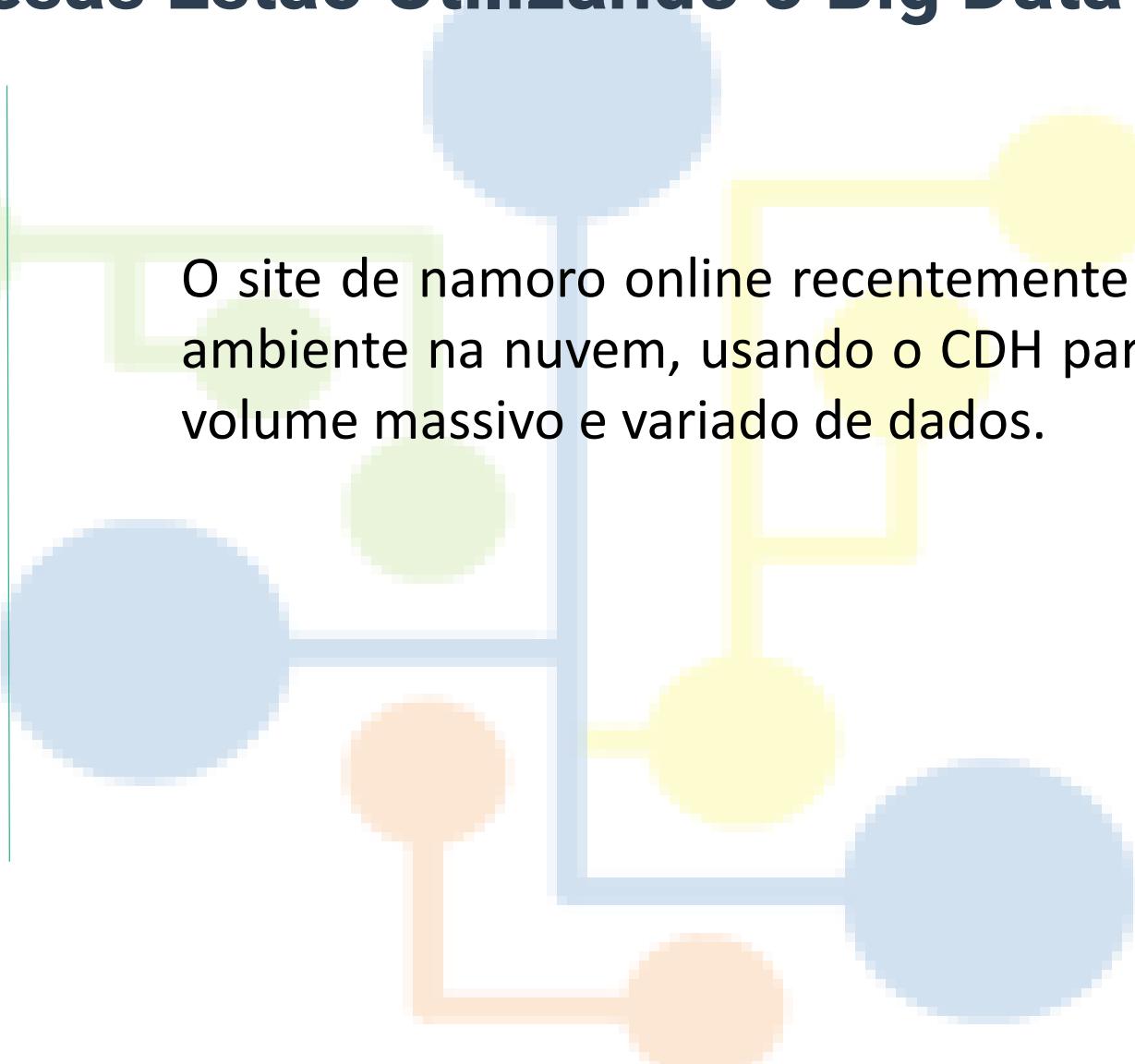
Entre outras coisas, ela colabora na determinação mais exata da probabilidade de um paciente estar com infecção em sua corrente sanguínea.



Como as Empresas Estão Utilizando o Big Data

The eHarmony logo, featuring the brand name in a teal sans-serif font with a registered trademark symbol.

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

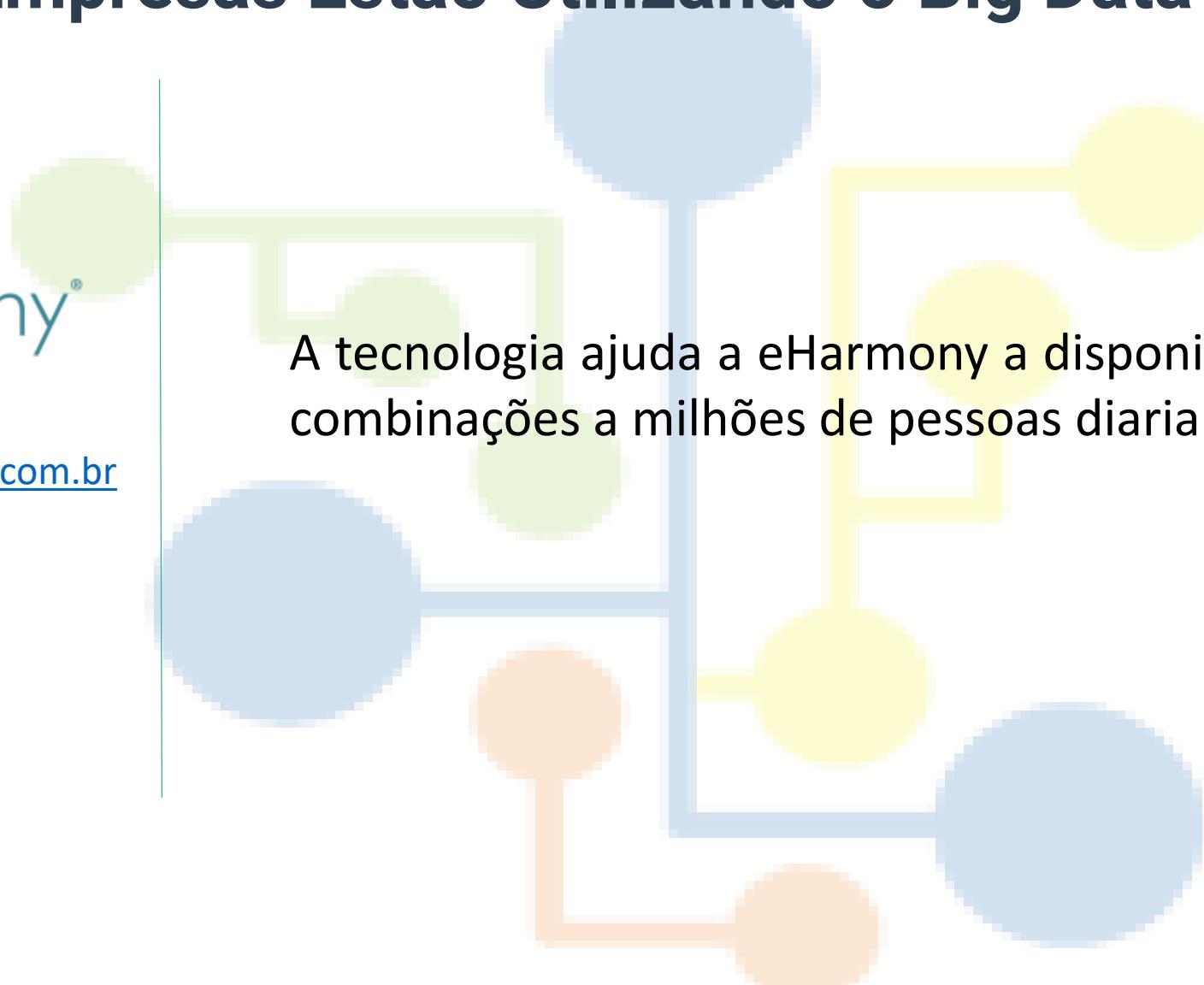


O site de namoro online recentemente atualizou seu ambiente na nuvem, usando o CDH para analisar um volume massivo e variado de dados.

Como as Empresas Estão Utilizando o Big Data

eHarmony®

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

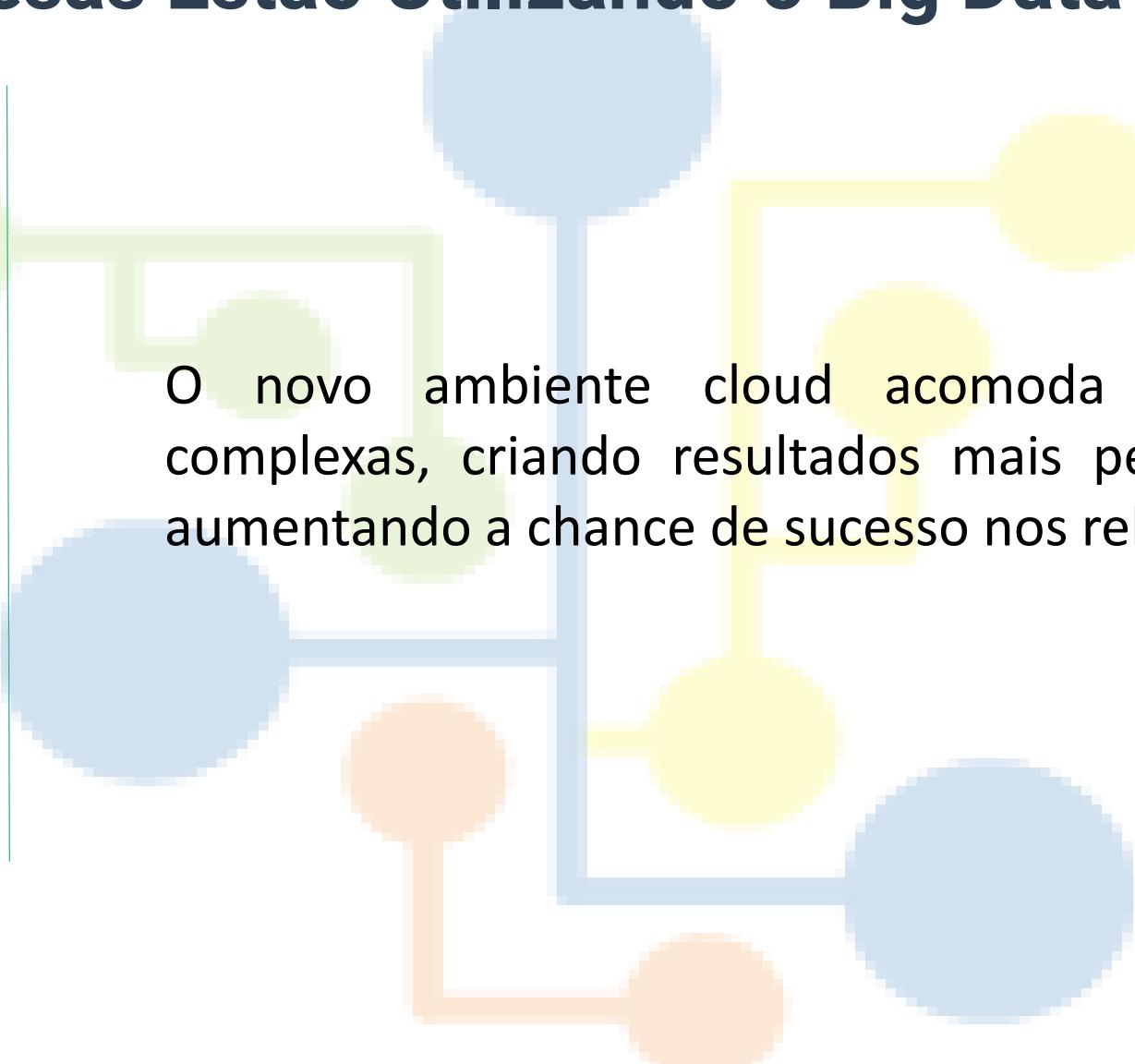


A tecnologia ajuda a eHarmony a disponibilizar novas combinações a milhões de pessoas diariamente.

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.



<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



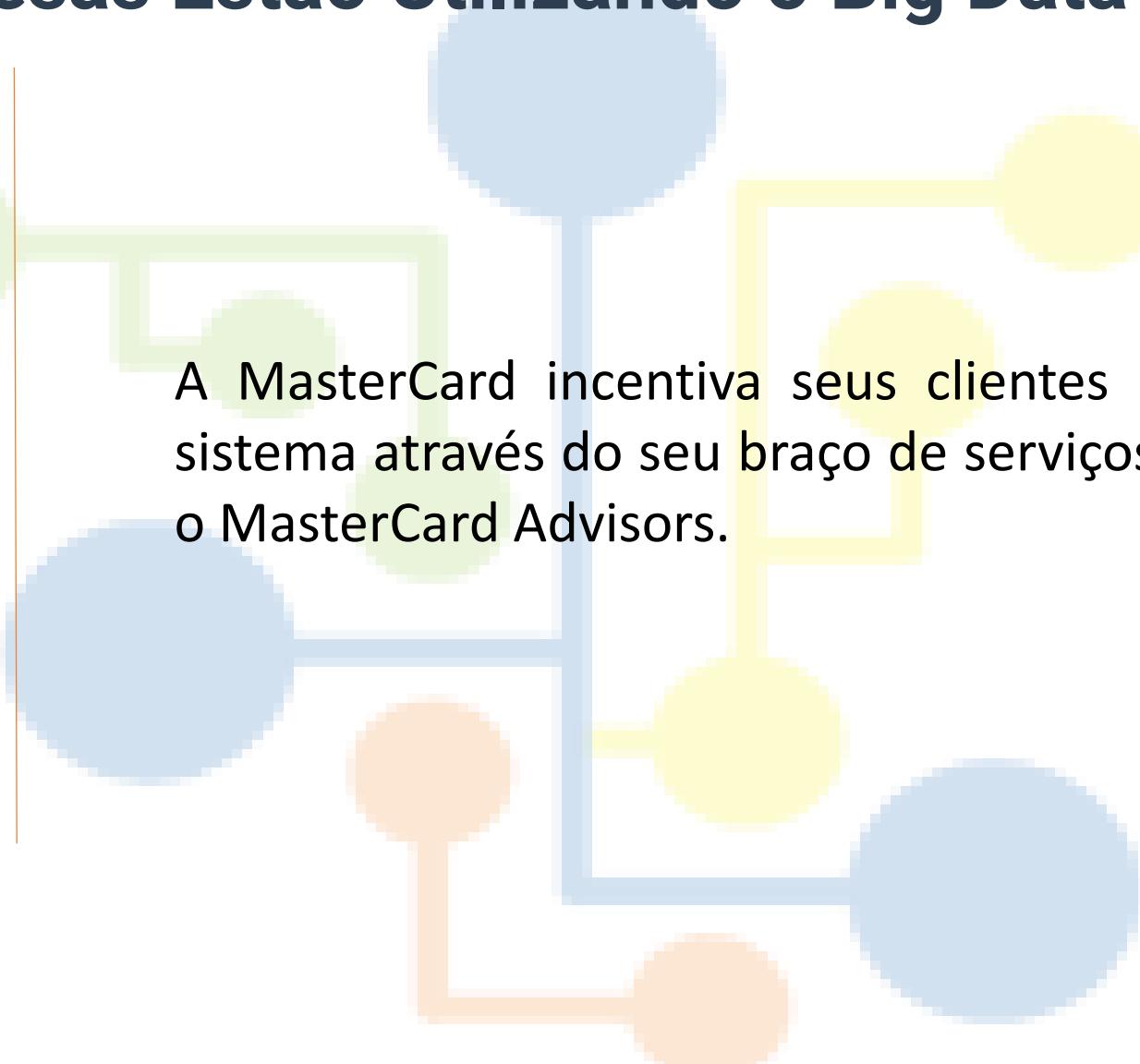
<http://www.mastercard.com.br>

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 decorative graphic consisting of several overlapping circles in various colors (red, orange, yellow, blue) connected by thin lines, forming a network-like pattern.

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



Estão Utile

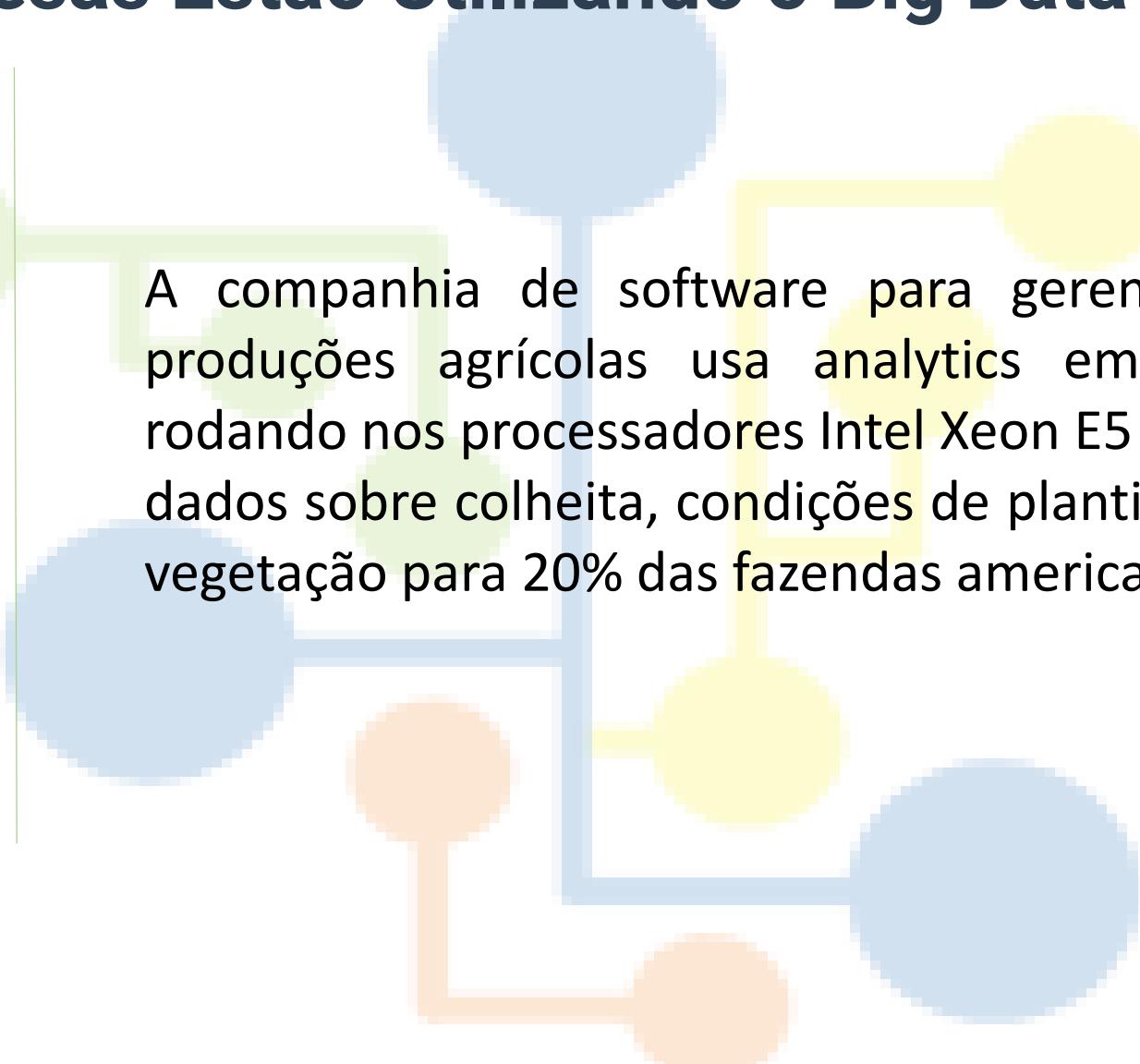
FarmLogs

<https://farmlogs.com>

Como as Empresas Estão Utilizando o Big Data

FarmLogs

<https://farmlogs.com>

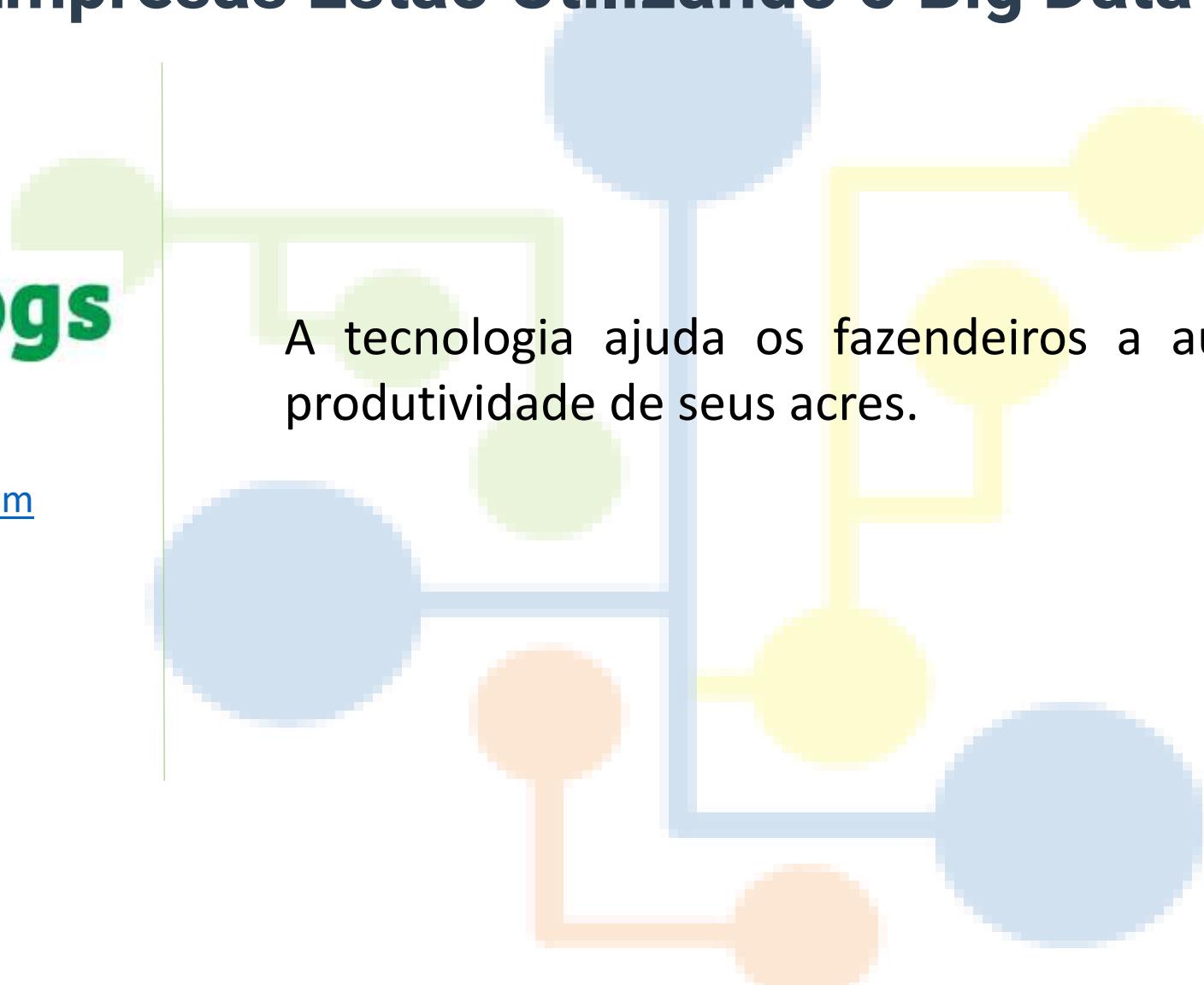


A companhia de software para gerenciamento de produções agrícolas usa analytics em tempo real rodando nos processadores Intel Xeon E5 para fornecer dados sobre colheita, condições de plantio e estado da vegetação para 20% das fazendas americanas.

Como as Empresas Estão Utilizando o Big Data

FarmLogs

<https://farmlogs.com>



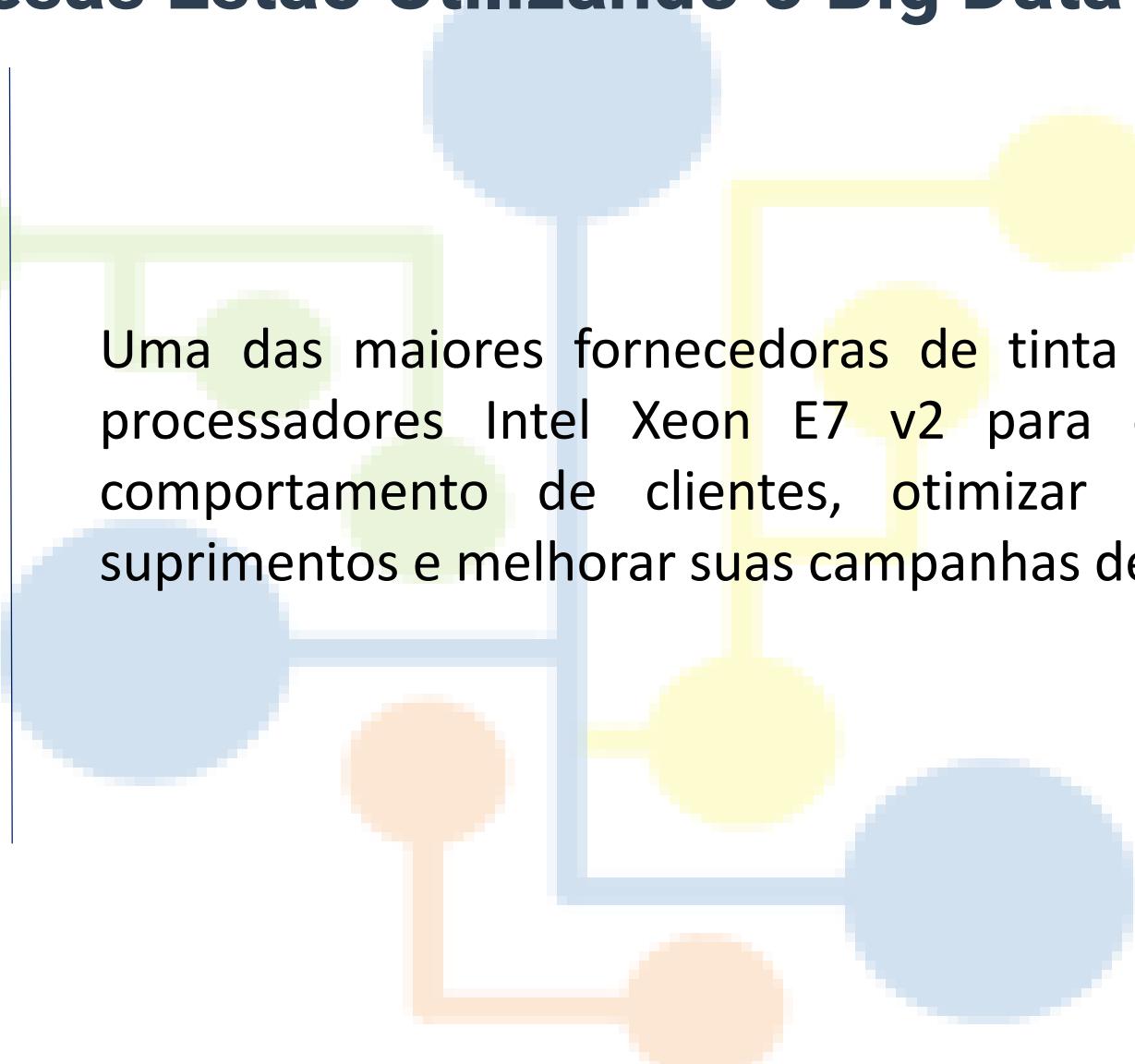
A tecnologia ajuda os fazendeiros a aumentarem a produtividade de seus acres.



Como as Empresas Estão Utilizando o Big Data



<http://www.nipponpaint.com>

An abstract graphic composed of several overlapping circles in light blue, yellow, and orange, connected by thin lines to form a network-like structure.

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



O Hadoop já é realidade!

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”



Data Science
Academy

Big Data Fundamentos 2.0

O futuro é aqui.



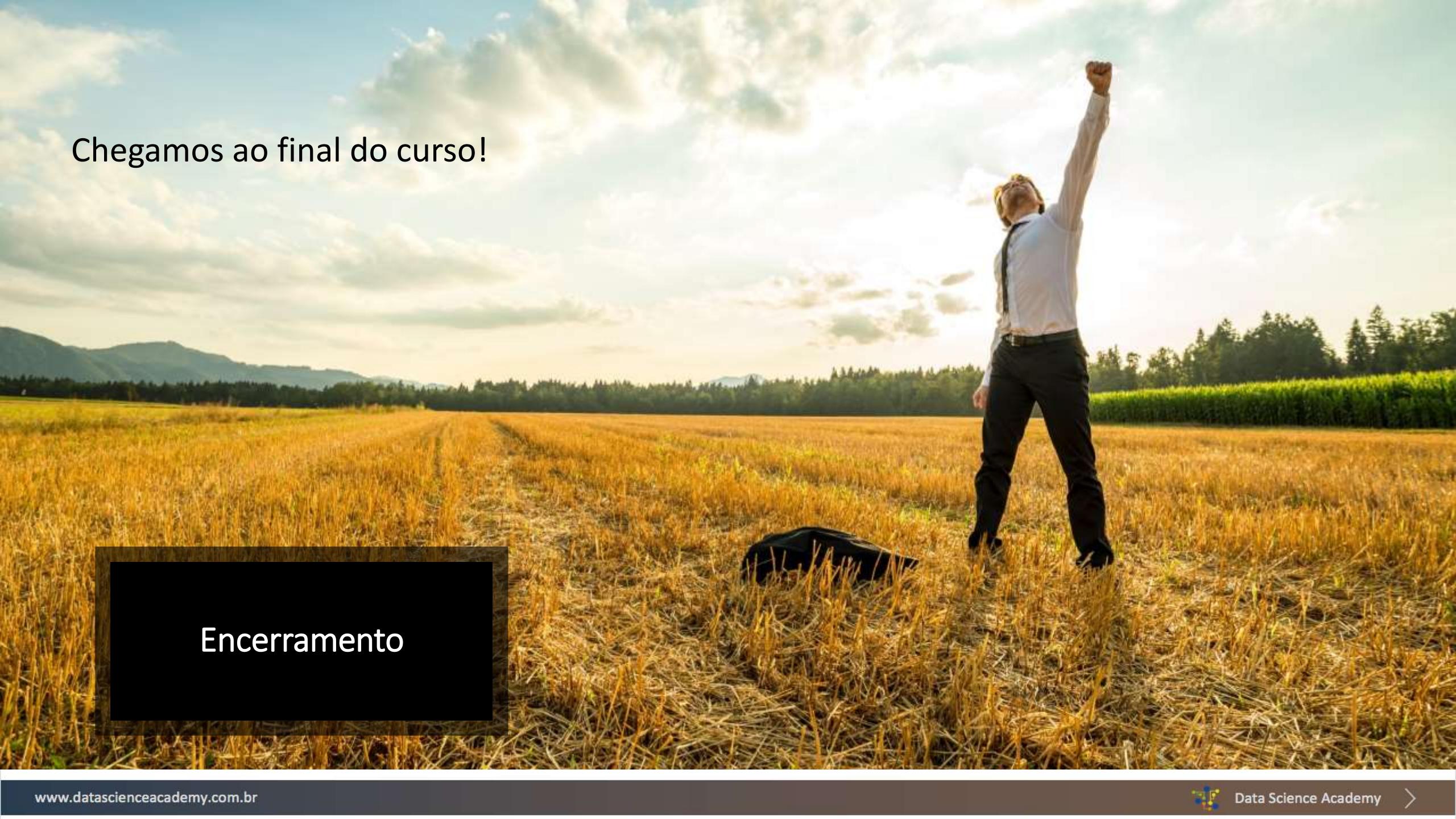
Conteúdo Programático

Seja Muito Bem-Vindo(a)!



Conteúdo do Curso

Visão geral de conceitos e definições que permitem uma compreensão clara do que é o universo do Big Data para que você possa avançar sua carreira nesta vibrante área.



Chegamos ao final do curso!

Encerramento

➤ Encerramento

O Brasil carece de profissionais capacitados em Big Data e que sejam capazes de construir e administrar um ambiente para coleta, armazenamento, limpeza, transformação e análise de dados.



Encerramento



E qual o Próximo Passo?

Encerramento

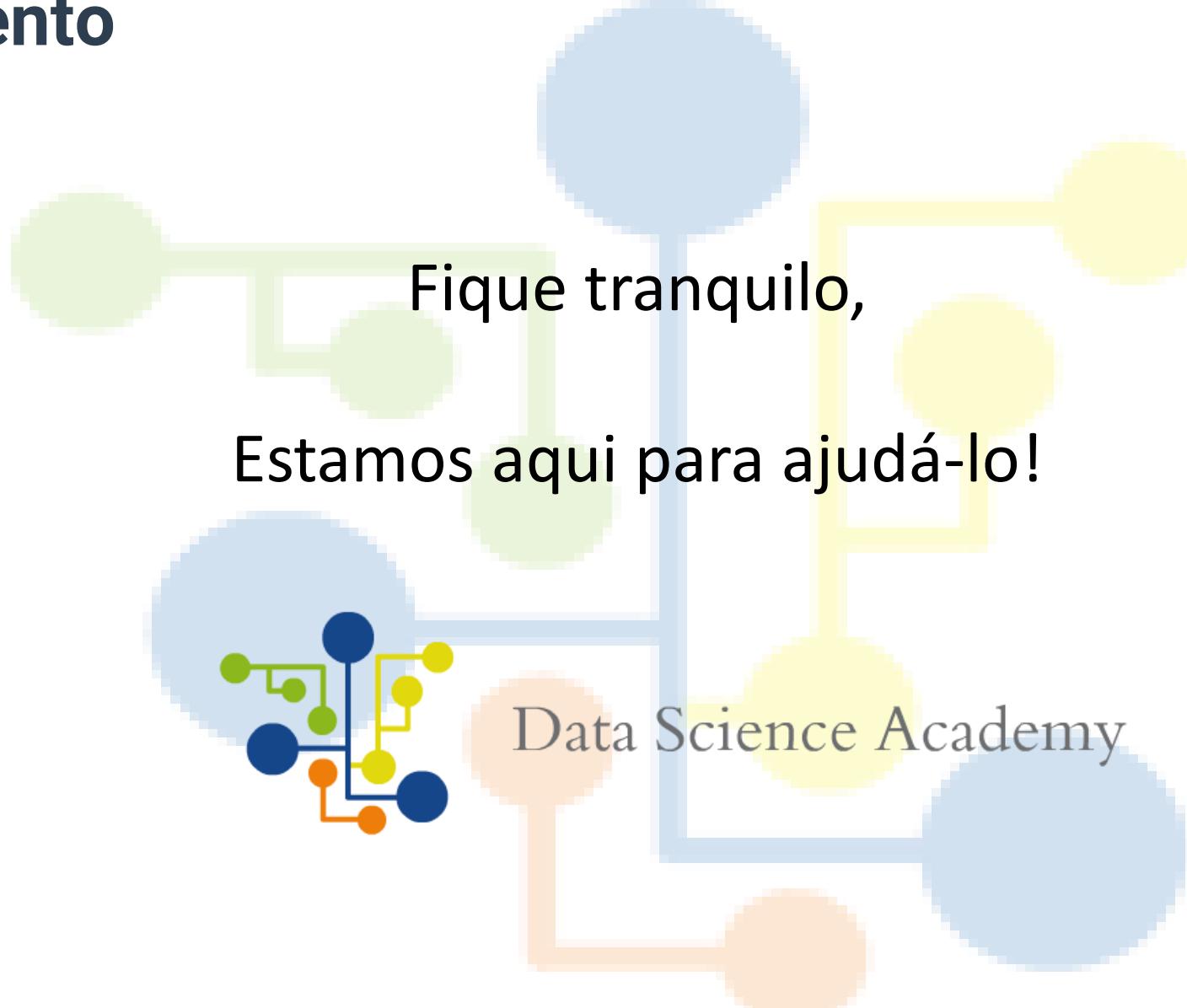
- Introdução à Ciência de Dados
- Python Fundamentos Para Análise de Dados
- Microsoft Power BI Para Data Science
- Formação Cientista de Dados
- Formação Inteligência Artificial
- Formação JAVA
- Formação Engenheiro de Dados
- Cursos Individuais

Encerramento

- Linguagem de programação para análise de dados (R, Python, Scala ou Java)
- Matemática e Estatística
- Algoritmos de Machine Learning
- Visualização de Dados
- Análise de dados distribuídos em Cluster
- Hadoop, Spark, Bancos de Dados NoSQL



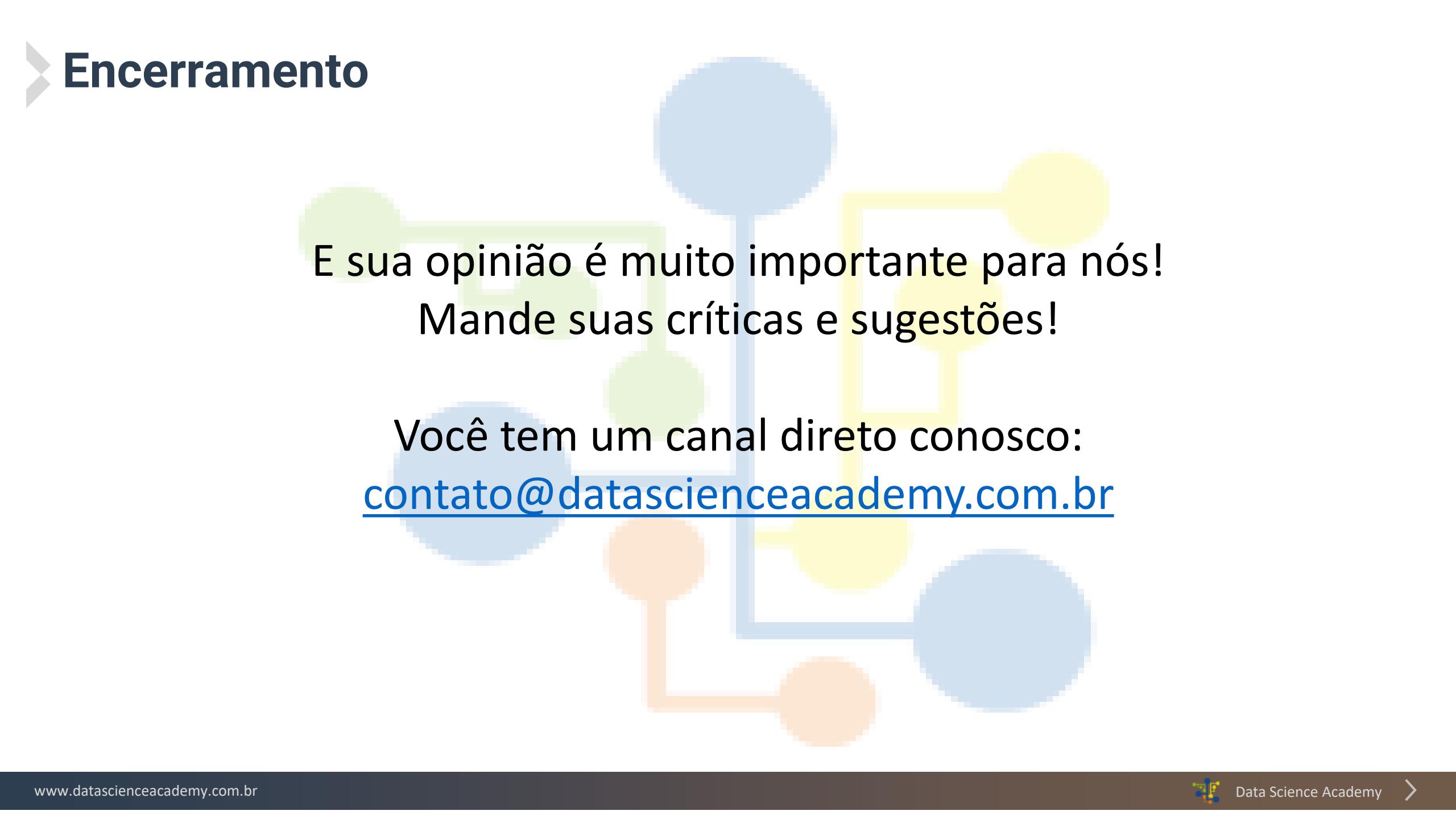
Encerramento



Fique tranquilo,
Estamos aqui para ajudá-lo!

Data Science Academy

Encerramento



E sua opinião é muito importante para nós!
Mande suas críticas e sugestões!

Você tem um canal direto conosco:
contato@datascienceacademy.com.br

Encerramento

Lembre-se de fazer a avaliação final para obter seu certificado.



Avaliação

Encerramento

Muito obrigado pela sua audiência neste curso e espero que tenhamos ajudado você na sua carreira!

