Com base na popularidade e usabilidade, listamos as dez ferramentas de código aberto a seguir como as melhores ferramentas de Big Data de código aberto em 2020.

**1. Hadoop**

O Apache Hadoop é a ferramenta mais importante e usada no setor de Big Data, com sua enorme capacidade de processamento de dados em larga escala. Essa é uma estrutura 100% de código aberto e é executada em hardware comum em um data center existente. Além disso, ele pode ser executado em uma infraestrutura de nuvem. O Hadoop consiste em quatro partes:

Sistema de arquivos distribuídos do Hadoop: Comumente conhecido como HDFS, é um sistema de arquivos distribuídos compatível com largura de banda de escala muito alta.

MapReduce: um modelo de programação para o processamento de Big Data.

YARN: é uma plataforma usada para gerenciar e agendar os recursos do Hadoop na infraestrutura do Hadoop.

Bibliotecas: Para ajudar outros módulos a trabalhar com o Hadoop.

Planejando construir uma carreira no Big Data Hadoop? Aqui estão os 20 termos mais importantes do Hadoop que você deve saber para se tornar um profissional do Hadoop.

### ****2. Apache Spark****

O Apache Spark é o próximo hype na indústria entre as ferramentas de Big Data. O ponto principal desta ferramenta de código aberto é que preenche as lacunas do Apache Hadoop em relação ao processamento de dados. Curiosamente, o Spark pode lidar com dados em lote e em tempo real. Como ele processa dados na memória, isso acontece muito mais rapidamente que o processamento em disco tradicional. Este é realmente um ponto positivo para os analistas que lidam com certos tipos de dados para obter resultados mais rápidos.

O Apache Spark é flexível para trabalhar com o HDFS e com outros armazenamentos de dados, por exemplo, com o OpenStack Swift ou o Apache Cassandra. Também é muito fácil executar o Spark em um único sistema local para facilitar o desenvolvimento e os testes.

O Spark Core é o coração do projeto e facilita muitas coisas como:

Transmissão de tarefas distribuídas;

Agendamento e

Funcionalidade de E / S.

O Spark é uma alternativa ao MapReduce do Hadoop, ele pode executar trabalhos 100 vezes mais rápido que o MapReduce do Hadoop. Se você quiser saber o motivo, leia o blog anterior sobre os 11 principais fatores que tornam o Apache Spark mais rápido.

### ****3. Apache Storm****

O Apache Storm é uma estrutura distribuída em tempo real para o processamento confiável do fluxo de dados ilimitado. Essa ferramenta suporta qualquer linguagem de programação e seus recursos exclusivos são:

Escalabilidade maciça;

Tolerância ao erro;

Abordagem “falhe rápido, reinicialização automática”;

O processo garantido de cada tupla;

Escrito em Clojure;

Executa na JVM;

Suporta topologia de gráfico acrílico direto (DAG);

Suporta vários idiomas e

Suporta protocolos como JSON.

As topologias de tempestade podem ser consideradas semelhantes ao trabalho do MapReduce. No entanto, no caso do Storm, é o processamento de dados de fluxo em tempo real, em vez de ser em lote. Com base na configuração da topologia, o planejador Storm distribui as cargas de trabalho para os nós. E ele pode interoperar com o HDFS do Hadoop através de adaptadores, se necessário, que é outro ponto que o torna útil como uma ferramenta de Big Data de código aberto.

### ****4. Cassandra****

O Apache Cassandra é um banco de dados de tipo distribuído para gerenciar um grande conjunto de dados nos servidores. Essa é uma das melhores ferramentas de Big Data que processa principalmente conjuntos de dados estruturados, que fornece serviço altamente disponível, sem ponto único de falha. Além disso, possui certos recursos que nenhum outro banco de dados relacional e qualquer banco de dados NoSQL podem fornecer. Esses recursos são:

Disponibilidade contínua como fonte de dados;

Desempenho escalável linear;

Operações simples;

Nos centros de dados, fácil distribuição destes;

Pontos de disponibilidade na nuvem;

Escalabilidade e

Atuação.

A arquitetura Apache Cassandra não segue a arquitetura mestre-escravo e todos os nós desempenham o mesmo papel. Ele pode lidar com vários usuários simultâneos nos data centers, portanto, adicionar um novo nó não importa no cluster existente, mesmo no período de atividade.

### ****5. RapidMiner****

O RapidMiner é uma plataforma de software para atividades de ciência de dados e fornece um ambiente integrado para:

Preparação de dados;

Aprendizado de máquina;

Mineração de texto;

Análise preditiva;

Aprendizagem profunda;

Desenvolvimento de aplicações e

Prototipagem.

Essa é uma das ferramentas úteis de Big Data que oferecem suporte a diferentes etapas do aprendizado de máquina, como:

Preparação de dados;

Visualização;

Análise preditiva;

Validação do modelo;

Otimização;

Modelagem estatística;

Avaliação e

Desdobramento, desenvolvimento.

O RapidMiner segue um modelo de cliente / servidor em que o último pode estar localizado no local ou em uma infraestrutura de nuvem. É escrito em Java e fornece uma GUI para projetar e executar fluxos de trabalho, além de poder fornecer 99% de uma solução analítica avançada.

### ****6. MongoDB****

O MongoDB é um banco de dados NoSQL de código aberto compatível com várias plataformas e com muitos recursos embutidos. É ideal para empresas que precisam de dados rápidos e em tempo real para tomar decisões instantâneas. E também para usuários que desejam experiências baseadas em dados. É executado na pilha de software MEAN, aplicativos NET e plataforma Java.

Alguns recursos notáveis ​​do MongoDB são:

Pode armazenar qualquer tipo de dados como inteiro, string, array, objeto, booleano, data etc.

Ele fornece flexibilidade na infraestrutura baseada em nuvem.

É flexível e particiona facilmente os dados entre os servidores em uma estrutura de nuvem.

O MongoDB usa esquemas dinâmicos, portanto você pode preparar dados rapidamente, sendo essa uma outra maneira de economizar custos.

### ****7. Ferramenta de Programação R****

Essa é uma das ferramentas de Big Data de código aberto amplamente usadas neste setor para análise estatística de dados. A parte mais positiva dessa ferramenta é, embora usada para análise estatística, como usuário, você não precisa ser um especialista neste quesito. O R possui sua própria biblioteca pública CRAN (Comprehensive R Archive Network), que consiste em mais de 9000 módulos e algoritmos para análise estatística de dados.

R pode ser executado no servidor Windows e Linux, bem como no servidor SQL etambém suporta Hadoop e Spark. Usando essa ferramenta, é possível trabalhar com dados discretos e experimentar um novo algoritmo analítico para análise. É uma linguagem portátil e sendo assim, um modelo R construído e testado em uma fonte de dados local pode ser facilmente implementado em outros servidores ou mesmo em um data lake Hadoop.

### ****8. Neo4j****

O Hadoop pode não ser uma escolha inteligente para todos os problemas relacionados ao Big Data. Por exemplo, quando você precisa lidar com um grande volume de dados de rede ou problemas relacionados a gráficos, como redes sociais ou padrão demográfico, um banco de dados de gráficos pode ser uma escolha perfeita.

O Neo4j é uma das ferramentas de Big Data que é amplamente usada no banco de dados de gráficos neste setor. Segue a estrutura fundamental do banco de dados de gráficos, que é a relação de dados interconectada dos nós. Ele mantém um padrão de valor-chave no armazenamento de dados.

Recursos notáveis ​​do Neo4j são:

Suporta transação ACID;

Alta disponibilidade;

Escalável e confiável;

Flexível, pois não precisa de um esquema ou tipo de dados para armazenar dados;

Pode integrar-se com outros bancos de dados e

Suporta linguagem de consulta para gráficos que é comumente conhecido como Cypher.

### ****9. Apache SAMOA****

O Apache SAMOA está entre as conhecidas ferramentas de Big Data usadas para algoritmos de streaming distribuídos para mineração de Big Data. Além de mineração de dados, ele também é usado para outras tarefas de aprendizado de máquina, como:

Classificação;

Agrupamento;

Regressão e

Abstração de programação para novos algoritmos.

É executado na parte superior dos DSPEs (mecanismos de processamento de fluxo distribuído). O Apache Samoa é uma arquitetura conectável e permite a execução em vários DSPEs que incluem:

Apache Storm;

Apache S4;

Apache Samza e

Apache Flink.

Devido às razões abaixo, Samoa tem imensa importância como a ferramenta de Big Data de código aberto do setor:

Você pode programar uma vez e executá-lo em qualquer lugar;

Sua infraestrutura existente é reutilizável, portanto você pode evitar a implantação de ciclos;

Sem tempo de inatividade do sistema e

Não há necessidade de processo complexo de backup ou atualização.

### ****10. HPCC****

O cluster de computação de alto desempenho (HPCC) é outro entre os melhores.

Ele é outra das melhores ferramentas de Big Data, considerada o concorrente do Hadoop neste mercado. É uma das ferramentas de Big Data de código aberto sob a licença Apache 2.0 e alguns dos seus principais recursos são:

Ajuda no processamento de dados paralelo;

Plataforma de computação de dados distribuídos de código aberto;

Segue arquitetura nada compartilhada;

Executa em hardware de commodity;

Vem com pacotes binários suportados para distribuições Linux ;

Oferece suporte ao gerenciamento completo do fluxo de trabalho de Big Data;

A plataforma inclui:

           Thor: para manipulação de dados orientada a lotes, seus links e análises e

           Roxie: para entrega e análise de dados em tempo real.

Implicitamente um mecanismo paralelo;

Mantém o encapsulamento de código e dados;

Extensível;

Altamente otimizado;

Ajuda a criar planos gráficos de execução e

Compila em C ++ e código de máquina nativo.

**Conclusão**

Para entrar no setor de Big Data é sempre bom começar com o Hadoop. Um treinamento de certificação no Hadoop associa muitas outras ferramentas de Big Data, como mencionado acima. Escolha qualquer um dos principais caminhos de certificação da Cloudera ou da Hortonworks e prepare-se para o mercado como profissional do Hadoop ou de Big Data.

## **O que é Big Data**

Big Data é o conjunto de informações presentes nos bancos de dados de servidores e empresas, que pode ser acessado e possui interligações entre si.

Ou seja, que está disponível na rede mundial de computadores e pode ser acessado remotamente.

Para ficar mais claro, o Youtube é um exemplo de BD, disponibilizando diversos vídeos presentes em um banco de dados disponível para usuários acessarem.

Outro exemplo é Wikipédia, com textos disponíveis em bancos de dados para consultas online. Porém, o **Big Data não se resume apenas a bancos de dados e informações online.**

Segundo a Lisa Arthur, contribuidora da revista Forbes, diversos CMO’s e CIO’s concordam que Big Data é todo e qualquer dado que pode ser coletado sobre uma empresa ou um assunto.

É uma coleção de informações de fontes tradicionais e digitais internos e externos à sua empresa, que representam uma fonte de descoberta e análise contínuas.

Apesar de os meios digitais reinarem quando o assunto é acúmulo de dados, as grandes empresas afirmam que não devem deixar de lado informações de compra e venda, registros financeiros e canais de interação não-digitais, como call centers e até telemarketing.

Tudo isso pode ser fonte de dados para ajudar a sua empresa a crescer.

## **Qual a história do Big Data**

Apesar de o termo Big Data ser relativamente novo, a premissa de reunir e armazenar informações é consideravelmente antiga.

Pense em bibliotecas, por exemplo!

Porém, o conceito ganhou uma abordagem nova no começo dos anos 2000, quando [Doug Laney](https://blogs.gartner.com/doug-laney/)escreveu um artigo que hoje é a definição mais conhecida de Big Data.

Doug separou a ideia em 5 V’s que vamos ver logo abaixo.

A crescente dos dados disponíveis e os modelos just in time de negócio, tornaram indispensáveis uma forma de **análise em tempo real de grandes quantidades de dados.**

Que é a grande diferença entre aplicar a ideia de BD na sua empresa, e armazenar informações sem qualquer tipo de organização ou análise.

## **O que são os 5 V’s do Big Data**

Atualmente, podemos dividir o Big Data em 5 V’s que formam a base para a implementação do conceito em qualquer empresa.

São eles: **Volume**, **Velocidade**, **Variedade**, **Variabilidade** e **Vínculo**.

### Volume

Volume é o ponto de partida para entender o BD.

Os 2,5 quintilhões de dados criados diariamente causa um volume de informações impressionante das mais diversas fontes.

Acompanhamento de leads e clientes, redes sociais, interações em sites e blogs, histórico de compra, cliques e muito, muito mais.

O volume de informações influencia em dois pontos principais: **armazenamento e análise.**

Com a expansão diária de servidores, ficou muito mais fácil armazenar esses dados, principalmente com a compressão de arquivos e possibilidade de disponibilizá-lo na nuvem.

Já a análise se torna cada vez mais simples, principalmente com o uso de ferramentas próprias para o Big Data.

### Velocidade

Dados são transferidos, armazenados e baixados em velocidades altíssimas!

E não estamos falando da velocidade da sua conexão com a internet, mas sim **como esse conteúdo é trabalhado, atualizado e expandido rapidamente.**

Cada vez mais fontes fornecem dados, o que torna necessário o gerenciamento dessas informações em tempo real, de forma ápida e segura.

### Variedade

Dados aparecem com todas as formas, cores e tamanhos. Podem ser planilhas, dados estruturados, ou somente documentos de texto, vídeos e imagens.

É preciso entender as variedades existentes e como cada uma deve ser analisada e armazenada.

### Variabilidade

A velocidade e variedade dos dados crescem constantemente, mas, adicionalmente, esses elementos mudam e não são necessariamente consistentes.

**Imagine as mídias sociais, como elas vão gerar informações de forma regular?**

Elas são totalmente dependentes das ações dos usuários e, mesmo parecendo, não ficamos 24 horas por dia conectados na web.

A ideia de variabilidade também pode ser aplicada a dados diários, sazonais e relativos a eventos específicos, o que gera um grande volume de informações e traz desafios no gerenciamento.

### Vínculo

Com um volume tão grande de dados, é muito provável que você perca o controle de tudo na hora que realmente precisar.

Um dos principais motivos disso é a dificuldade de conectar e transformar informações através de diferentes plataformas.

Portanto, é preciso conectar e correlacionar os elementos, criar hierarquias e ligações múltiplas para os dados, ou seja, criar vínculos entre eles.

## **Por que Big Data foi criado**

O volume incrivelmente alto de informações gerados diariamente — e acumulado de outros anos — passou a ser visto como uma fonte de insights e não só um apanhado de dados.

Logo, foi preciso pensar em um modelo de análise que ajude as empresas a encontrar ideias valiosas no meio de tantos bytes.

O Big Data ajuda empresas a encontrarem oportunidades não só onde elas estão claras, mas em **correlações e cruzamento de dados complexos**, na decupagem de dados estruturados, não estruturados e multi-estruturados.

## **Como interpretar dados**

É muito provável que você já entendeu a importância dos **dados**para Big Data. Afinal, já falamos sobre eles diversas vezes.

Existe mais de um tipo de dado.

Basicamente, qualquer informação pode ser definida assim, mas existem diferentes formatos que podem facilitar, dificultar ou direcionar a análise feita pelas empresas.

Os dados são divididos em dois tipos: estruturados e não estruturados. Vamos conhecer um pouco mais sobre cada um deles?

### Dados não estruturados

Com o passar dos anos, os bancos de dados passaram a interpretar com mais facilidade dados.

Isso aconteceu muito pelo fato de esses bancos armazenarem informações similares, como planilhas, cálculos, funções e números, o que facilita a forma como esse conteúdo é analisado.

Porém, novos formatos chegaram e eles não são tão facilmente interpretados como antes.

Eles são chamados de **dados não estruturados**, informações que **não estão organizadas ou que não são facilmente entendidas pelos bancos de dados tradicionais** e formatos conhecidos de dados.

Geralmente, esses elementos são, predominantemente, textos. Metadata de blogs, imagens, e tuítes são exemplos de dados não estruturados.

### Dados estruturados (ou multi-estruturados)

Em contrapartida, existem dados que são facilmente reconhecidos pelos bancos e facilitam a análise e processamento. Eles são chamados de**dados estruturados** (ou multi-estruturados).

Eles são derivados das interações entre pessoas e máquinas, como aplicações na web e mídias sociais.

Um bom exemplo são dados resultantes do comportamento dos usuários na web — também conhecidos como **logs**.

É uma mistura de texto, imagens e dados estruturados como formulários ou informações transacionais.

O avanço digital transformou muitos desses dados, trazendo ainda mais formatos para os considerados estruturados.

Com a constante evolução das relações entre marcas, consumidores e plataformas, a tendência é que essas informações continuem evoluindo e se alterando.

## **Quais são as aplicações do Big Data em diferentes setores?**

Atualmente, o que faz o Big Data ser tão bem-sucedido é a sua aplicação em inúmeros contextos. Podemos ver usos da análise de grandes volumes dados para efeitos diversos, muitos deles vinculados à geração de melhores resultados de lucratividade nas empresas.

### Comportamento do usuário

O comportamento do usuário pode ser rastreado e coletado com o apoio do Big Data. A equipe pode dispor os dados de forma organizada em [dashboards](https://rockcontent.com/blog/dashboard/) para visualização e compartilhamento. A partir do uso de sistemas de analytics, é possível **intuir determinadas ações do usuário de acordo com esse comportamento prévio.**

Com base em históricos de compras, em interações e em sessões de atendimento, o sistema pode predizer, por exemplo, a possibilidade de um cliente pedir o cancelamento dos serviços contratados.

No Marketing, o Big Data pode ser o combustível para a análise preditiva, que busca prever como um cliente vai reagir a determinadas táticas. É viável, por exemplo, amplificar sua mensagem e obter anúncios otimizados para gerar o máximo de conversão e de vendas ao estudar quais formas de comunicação ressoam melhor com o público.

Além de prever, é possível personalizar a abordagem para afunilar a comunicação.**Desse modo, a empresa pode satisfazer e fidelizar os seus consumidores, para obter melhores resultados.**

### Logística

Na logística, o uso de análise de dados ajuda a gerar eficiência na definição das melhores rotas para levar um produto de A a B. É possível otimizar o transporte e gerenciar melhor os riscos nessa frente. Da mesma forma, a empresa consegue antecipar demandas para preparar os seus estoques.

### Saúde

Na saúde, existe uma longa frente tecnológica que adota soluções de Big Data. Uma delas busca prever doenças nas pessoas com base em uma série de dados históricos, de monitoramento. Além disso, é possível detectar as chances de algum quadro clínico ocorrer de acordo com padrões de outras doenças envolvendo características de cada paciente.

Nesse nicho, é importante destacar a presença da internet das coisas, IoT. Os dispositivos da IoT conseguem monitorar e acompanhar as condições de um paciente, gerando **uma quantidade significativa de dados**que podem ser analisados.

Portanto, a IoT produz informações que devem ser enviados a sistemas analíticos que ajudam no tratamento e controle de doenças.

### Detecção de fraudes

Com dados diversos acerca de operações financeiras e possíveis ocorrências de fraudes, torna-se viável detectar esses riscos e antecipá-los de forma proativa.

Dessa forma, quaisquer sinais que indicam a possibilidade de uma atividade fraudulenta já levantam um alerta para a empresa, que pode agir de **forma preventiva** para combater as consequências.

### Sistemas de recomendação

Outro uso do Big Data associado ao varejo ou ao marketing é em sistemas de recomendação.

Nesse caso, são coletados dados variados acerca dos usuários e de suas preferências, **de modo a gerar recomendações para pessoas semelhantes.**

O sistema analítico chega a uma probabilidade para indicar algo a alguém, que pode ser um produto ou um filme a ser assistido. É o mecanismo utilizado pela Netflix nas telas iniciais de cada usuário!

### Otimização de preços

Podemos mencionar também o uso de sistemas que processam e analisam dados para chegar a uma otimização dinâmica de preços.

Assim, em vez de somente definir valores estáticos, o sistema consegue analisar questões da concorrência, preferências do usuário, contexto socioeconômico e outras variáveis para mudar o preço sempre que for necessário. Essas ações são feitas para maximizar as vendas e a lucratividade em tempo real.

### Assistentes virtuais

Os assistentes virtuais são um ótimo exemplo de como transformar os dados em aplicações inteligentes. Nesse sentido, são coletados dados de interações com o usuário para análise de sistemas analíticos que geram um resultado capaz de dialogar e interagir de maneira espontânea com os clientes.

É viável ainda personalizar a abordagem de comunicação de acordo com a conversa e transmitir emoções variadas, que sejam coerentes com o tom de voz da marca e o perfil do cliente ideal.

### Manutenção preditiva

Também associada com a IoT, a manutenção preditiva se destaca por oferecer novamente uma capacidade preventiva para as empresas.

Com a análise dos sensores, é possível coletar uma quantidade enorme de dados para processamento. Uma vez que esses dados são coletados, **a empresa é capaz de fazer reparos inteligentes antes que determinada máquina pare de funcionar.**

## **Qual a relação entre Big Data, Machine Learning e Inteligência Artificial?**

As aplicações que mencionamos anteriormente combinam Big Data, [Machine Learning e Inteligência Artificial](https://rockcontent.com/br/blog/inteligencia-artificial-machine-learning/). É preciso, portanto, saber a diferença principal entre esses conceitos e a relação entre eles.

Basicamente, o que chamamos de Big Data (dados estruturados ou não, em grande volume, gerados rapidamente e com formatos variados) **é a base para os treinamentos de modelos de Machine Learning (ML)**.

Todo sistema de ML precisa de um treinamento, seja ele supervisionado, não supervisionado ou por reforço. Dessa forma, é preciso uma etapa anterior ao seu funcionamento para “ensinar” o modelo.

Esse treinamento é feito com base em dados diversos. Quanto maior a complexidade do nosso problema de ML, maior deve ser essa base. Com os sistemas de ML modernos, **temos a vantagem de poder coletar dados em diferentes formatos, já que os algoritmos aceitam uma variedade maior de opções.**

Assim, podemos coletar uma tabela estruturada do Excel ou tweets do Twitter tranquilamente. Esses dados são logo preparados, convertidos e transformados para servir como base para o treinamento do modelo inteligente.

Ou seja, com as bibliotecas e as ferramentas de manipulação, conseguimos lidar com dados de todos os tipos, não somente os relacionais e estruturados. Por isso, o ML gerencia perfeitamente os dados que chamamos de Big Data.

### Deep learning

Dentro do conceito de Machine Learning, temos um tipo específico de algoritmo chamado de redes neurais. Além disso, existe um meio mais complexo de rede neural que é chamado de deep learning, que é caracterizado por precisar de ainda mais dados para o seu treinamento. Por isso, se alimenta também do Big Data.

### Relação com a inteligência artificial

**O machine learning, por sua vez, é um subcampo da Inteligência Artificial (IA)**, que praticamente monopolizou o debate sobre IA nos últimos anos. Além disso, é um dos meios mais modernos e bem-sucedidos da IA.

### MLOps

Como uma das características do Big Data é a sua natureza constantemente mutável, temos que abordar também a questão do MLOps e do pipeline de dados.

Nesse tipo de projeto, o modelo é retreinado sempre que houver uma atualização nos dados, de modo a manter a acurácia das predições e a consistência dos resultados. **Isso é feito de forma automática, acompanhando a volatilidade das fontes de entrada.**

## **Big Data e Marketing: como se conectam?**

Não se engane, marketing (e vendas) podem se beneficiar do Big Data tanto quanto qualquer outro setor. Talvez até mais do que qualquer outro!

Isso acontece porque empresas que baseiam seu **marketing em dado**s — também conhecido como **data-driven marketing** — dominam o mercado, sem sombra de dúvidas.

Os times de marketing mais eficientes são aqueles que colhem as informações recolhidas através das leads, do comportamento dos usuários e das experiências do time, e as transformam em ações que ajudam a alcançar resultados.

Separamos 5 aspectos do marketing que são favorecidos com o uso de Big Data. Confira!

### Interpretação de dados

**Empresas que tomam decisões a partir de dados são mais felizes.**

É verdade!

Elas são mais felizes por que essas resoluções são baseadas em análises, dados e relatos que ajudam na sua eficiência.

**Uma decisão tem mais chances de trazer os resultados esperados quando é baseada em dados** que comprovam sua eficiência projetada.

Uma [pesquisa da KPMG Capital](https://www.decisionanalyticsgroup.com/getting-past-the-data/) mostrou que 99% dos entrevistados acreditam no Big Data como importante parte da tomada de decisão.

Porém, 85% deles alegam que têm dificuldade em interpretar os dados e apenas 25% aplicam os insights obtidos através da análise.

Logo, é preciso de um profissional capacitado para interpretar devidamente as informações e ajudar a empresa a tomar as decisões certas.

Para essa interpretação, as ferramentas são tão importantes quanto os profissionais.

Em sua maioria, são plataformas de organização e decupagem de dados, que facilitam a inferência deles em um futuro próximo.

### Análises de mercado

A intuição e experiência dos profissionais direcionou empresas ao sucesso.

Mas nem sempre o caminho foi tão tranquilo assim. Muitos erros, testes e dinheiro foram gastos no processo.

Muito disso aconteceu porque não é possível adivinhar o que o seu público quer e o Big Data é uma mão na roda nessa hora.

As personas — perfis semificcionais — são indispensáveis para uma estratégia de marketing digital de qualidade, e para criar boas personas você precisa ter informações sólidas sobre o que elas gostam, onde estão presentes e qual conteúdo desejam consumir.

Nessa hora, encontrar padrões de comportamento entre os clientes da sua empresa e como eles atuam dentro do mercado seria bem interessante, não é mesmo?

Análises descritivas de eventos, correlações e **ideias geradas a partir dos dados** são as fontes de conhecimento que o Big Data fornece para a sua estratégia de marketing.

### Experiência de compra

A tomada de decisão de compra pode ser um grande mistério.

O que leva uma pessoa a escolher a sua empresa?

Os motivos podem ser vários e se você está envolvido no mundo do [trade marketing](https://rockcontent.com/br/blog/o-que-e-trade-marketing/), venda direta ou marketing multinível, Big Data pode ser a sua melhor amiga — e talvez você nem saiba.

Analisar esses dados ajuda o seu negócio a encontrar os melhores meios de distribuição para os produtos, técnicas de PDV (ponto de venda) e fornecer a experiência de compra que o cliente espera.

Em mercados multinível, existe uma demanda grande por personalização por parte do cliente e você precisa de conhecimento estratégico para atender as necessidades da sua persona.

### Mais fontes de dados para a sua empresa

No início do Big Data, as fontes de dados estavam limitadas a informações geradas por algumas empresas em alguns programas.

Claramente, nós vimos como o volume de referências aumentou exponencialmente ao longo dos anos. **E isso não é nem um pouco negativo.**

Claro, agora, é preciso ser um especialista para conseguir as estatísticas realmente relevantes a partir dos dados disponíveis, mas, ao mesmo tempo, a variedade desses elementos é cada vez maior.

[Twitter](https://rockcontent.com/br/blog/ganhar-seguidores-no-twitter/), [Facebook](https://rockcontent.com/br/blog/facebook/), [Instagram](https://rockcontent.com/br/blog/instagram-stories/)e diversos aplicativos fornecem dados diariamente.

Um varejista pode utilizar dados meteorológicos para comprar os produtos certos na hora certa.

Um organizador de eventos pode escolher a melhor data para uma ação através da previsão do tempo, e assim por diante.

**Todas as fontes de informação ajudam a sua empresa a tomar as melhores decisões, encontrar as melhores soluções para seus clientes e impulsionar as vendas.**

Os logs da web foram expandidos e se tornaram o conteúdo das redes sociais, dados de ferramentas de BI (Business Intelligence), relatórios, indicadores macroeconômicos e muito mais.

Cada vez mais fontes fornece informações relevantes para empresas.

### Automação de marketing

A nossa velha conhecida automação de marketing não é só indispensável para estratégias de marketing de conteúdo e [inbound marketing](https://rockcontent.com/br/blog/o-que-e-inbound-marketing/) de sucesso.

Elas são mais fontes de dados e também fazem análises interessantes para as empresas.

Se pensarmos bem, as plataformas de automação armazenam informações sobre o usuário, características de comportamento, e até métricas de rendimento de campanhas e ações.

Logo, é mais uma parte do Big Data altamente relevante para a sua empresa.

Além disso, são ferramentas com muitas integrações e que te ajudam a ter uma visão cada vez mais ampla do material que você tem disponível para análise.

## **O que é Big Data Analytics**

Com certeza você já percebeu que é preciso analisar para crescer. Aliás, foi esse fato que motivou tanto o crescimento do conceito de Big Data.

Logo, o lado analista não pode ser deixado de lado.

**Big Data é o conjunto dos dados dentro dos bancos, já Big Data Analytics é colocar a mão na massa e encontrar os insights** sobre tendências do mercado, comportamento dos consumidores e suas expectativas.

O que é o analytics faz é pegar todas as fontes de dados que nós já vimos anteriormente — a matéria-prima de tudo — e criar as correlações, inferências e vínculos entre as informações.

Assim, é possível encontrar as métricas relevantes para a sua empresa, usar números para tomar decisões e criar estratégias mais eficientes.

### Identificando padrões

Padrões são bem importantes.

Os padrões da sua audiência te ajudam a montar a sua persona, usuários que se comportam de forma similar e se tornam clientes dão insights sobre as ações que você deve executar para conquistar mais clientes desse tipo.

O [remarketing](https://rockcontent.com/br/blog/remarketing/) é um bom exemplo desses padrões e análises. Depois de se interessar por um livro, quantas ofertas semelhantes você já recebeu após isso?

Essa é uma das aplicações do conhecimento obtido através desse diagnóstico.

### Concorrência

Conhecer e entender o que a sua concorrência está fazendo é bem importante para o sucesso do seu negócio.

Não necessariamente você vai atacar todos os seus competidores, mas a verdadeira vantagem está em aprender com os erros deles.

**A análise de dados é uma excelente ferramenta para encontrar os melhores caminhos à serem percorridos**, tomando como base as ações dos seus concorrentes que não apresentaram os resultados esperados.

Grande parte desses insights vem do comportamento do seu público e te ajuda a criar promoções específicas para o seu mercado.

### Big data analytics no marketing e vendas

Métricas são importantes para ambos os times.

O marketing precisa entender quais ações estão gerando mais resultados — gerando mais leads, quais tem o melhor [ROI](https://rockcontent.com/br/blog/roi-do-marketing-de-conteudo/) quais atraem mais tráfego — e o time de vendas precisa encontrar as ações e processos que convertem mais leads em clientes.

A análise já faz parte do dia a dia da maioria desses times e se não faz parte dos times da sua empresa, é hora de começar!

**Big Data ajuda as empresas a entender padrões de comportamento e de respostas às ações realizadas.**

Muitas empresas usam essa análise para reduzir o seu[Custo de Aquisição por Cliente (CAC)](https://rockcontent.com/br/blog/custo-de-aquisicao-de-clientes/), customer [lifetime value](https://rockcontent.com/br/blog/lifetime-value/) (CLTV), otimizar preços, descontos e muito, muito mais.

O segredo das empresas de sucesso é tornar o Big Data Analytics uma parte ativa do cotidiano  dos times.

O Walmart critou o [Walmart Labs,](https://www.walmartlabs.com/?utm_source=datafloq&utm_medium=ref&utm_campaign=datafloq) que ajuda a empresa a criar ferramentas de relacionamento com o público através dos dados colhidos sobre os clientes.

O MIT utilizou o MIT Media Lab para mapear quantas pessoas se encontravam no maior centro de compras dos EUA em um determinado momento, durante a Black Friday.

A Nike utiliza o seu app de corrida para incentivar seus usuários a praticarem esportes e compartilharem nas redes sociais, e ainda recolhe diversas informações relevantes para o negócio.

Bem interessante, não é?

## **Conclusão**

A ideia de Big Data procura transformar tudo o que as empresas coletam no dia a dia de seus clientes — em forma de dados e informações — em oportunidades valiosas para o negócio.

Afinal, com tanto conteúdo disponível, tanto dos seus próprios clientes quanto de visitantes e desconhecidos, não aproveitar essa chance de melhorar a sua empresa realmente não é a melhor opção.

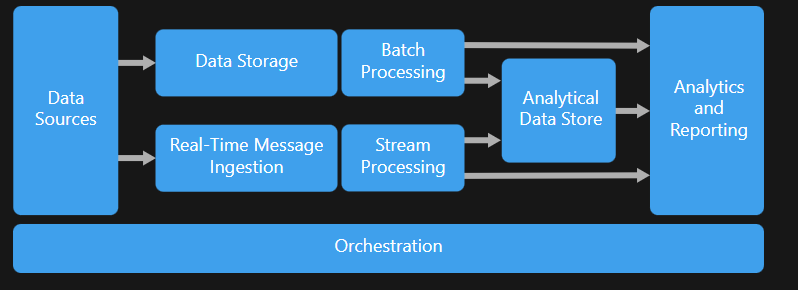
O poder de analisar cada vez mais criou oportunidades para mais profissionais analíticos no mercado, mas que também precisam conhecer o segmento do negócio e, ainda mais importante, entender como aplicar o que foi aprendido na empresa.

**Você não precisa fazer parte de uma grande empresa para aplicar a ideia do Big Data.**

Ferramentas gratuitas como o Google Analytics e até mesmo outras de gerenciamento de redes sociais, email marketing e automação de marketing podem render insights interessantes.

Tanto que um dos principais pontos de qualquer estratégia de análise de qualidade é exportar os relatórios certos.

Uma arquitetura de Big Data foi projetada para lidar com ingestão, processamento e análise de dados grandes ou complexos demais para sistemas de banco de dados tradicionais.



Soluções de Big Data normalmente envolvem um ou mais dos seguintes tipos de carga de trabalho:

Processamento em lote de fontes Big Data em repouso.

Processamento em tempo real de Big Data em movimento.

Exploração interativa de Big Data.

Análise preditiva e machine learning.

A maioria das arquiteturas de Big Data inclui alguns ou todos os seguintes componentes:

**Fontes de dados**: todas as soluções de Big Data começam com uma ou mais fontes de dados. Os exemplos incluem:

Armazenamentos de dados de aplicativo, como bancos de dados relacionais.

Arquivos estáticos produzidos por aplicativos, como arquivos de log do servidor Web.

Fontes de dados em tempo real, como dispositivos IoT.

**Armazenamento de dados**: dados de operações de processamento em lote normalmente são armazenados em um repositório de arquivos distribuído que pode conter amplos volumes de arquivos grandes em vários formatos. Esse tipo de repositório geralmente é chamado *data lake*. As opções para implementar esse armazenamento incluem contêineres de blobs ou Azure Data Lake Store no Armazenamento do Azure.

**Processamento em lote**: como os conjuntos de dados são muito grandes, geralmente uma solução de Big Data deve processar arquivos de dados usando trabalhos de lote de execução longa para filtrar, agregar e preparar os dados para análise. Normalmente, esses trabalhos envolvem ler arquivos de origem, processá-los e gravar a saída para novos arquivos. Opções incluem executar trabalhos de U-SQL no Azure Data Lake Analytics, usar trabalhos Hive, Pig ou de Mapear/Reduzir personalizados em um cluster HDInsight Hadoop ou usar programas de Java, Scala ou Python em um cluster HDInsight Spark.

**Ingestão de mensagens em tempo real**: se a solução inclui fontes em tempo real, a arquitetura deve incluir uma maneira de capturar e armazenar mensagens em tempo real para processamento de fluxo. Isso pode ser um armazenamento de dados simples, em que as mensagens de entrada são removidas para uma pasta para processamento. No entanto, muitas soluções precisam de um repositório de ingestão de mensagens para atuar como buffer de mensagens e dar suporte a processamento de expansão, entrega confiável e outras semânticas de enfileiramento de mensagem. Opções incluem Hubs de Eventos do Azure, Hubs de IoT do Azure e Kafka.

**Processamento de fluxo**: depois de capturar mensagens em tempo real, a solução deve processá-las filtrando, agregando e preparando os dados para análise. Os dados de fluxo processados são gravados em um coletor de saída. O Azure Stream Analytics oferece um serviço de processamento de fluxo gerenciado baseado em consultas SQL em execução perpétua que operam em fluxos não associados. Você também pode usar tecnologias de streaming Apache de software livre, como Storm e Spark Streaming em um cluster HDInsight.

**Armazenamento de dados analíticos**: muitas soluções de Big Data preparam dados para análise e então veiculam os dados processados em um formato estruturado que pode ser consultado usando ferramentas analíticas. O armazenamento de dados analíticos usado para atender a essas consultas pode ser um data warehouse relacional estilo Kimball, como visto na maioria das soluções de BI (business intelligence) tradicionais. Como alternativa, os dados podem ser apresentados por meio de uma tecnologia NoSQL de baixa latência, como HBase ou um banco de dados Hive interativo que oferece uma abstração de metadados sobre arquivos de dados no armazenamento de dados distribuído. O Azure Synapse Analytics fornece um serviço gerenciado para armazenamento de dados em larga escala baseado em nuvem. O HDInsight dá suporte a Hive interativo, HBase e Spark SQL, que também pode ser usado para veicular dados para análise.

**Análise e relatório**: a meta da maioria das soluções de Big Data é gerar insights sobre os dados por meio de análise e relatórios. Para capacitar os usuários a analisar os dados, a arquitetura pode incluir uma camada de modelagem de dados, como um cubo OLAP multidimensional ou um modelo de dados tabular no Azure Analysis Services. Também pode dar suporte a business intelligence de autoatendimento, usando as tecnologias de modelagem e visualização do Microsoft Power BI ou do Microsoft Excel. Análise e relatórios também podem assumir a forma de exploração de dados interativos por cientistas de dados ou analistas de dados. Para esses cenários, muitos serviços do Azure dão suporte a blocos de anotações analíticos, como Jupyter, permitindo que esses usuários aproveitem suas habilidades existentes com Python ou R. Para exploração de dados em larga escala, você pode usar o Microsoft R Server, seja no modo autônomo ou com Spark.

**Orquestração**: a maioria das soluções de Big Data consiste em operações de processamento de dados repetidos, encapsuladas em fluxos de trabalho, que transformam dados de origem, movem dados entre várias origens e coletores, carregam os dados processados em um armazenamento de dados analíticos ou efetuam o push dos resultados diretamente para um relatório ou painel. Para automatizar esses fluxos de trabalho, você pode usar uma tecnologia de orquestração, como Azure Data Factory ou Apache Oozie e Sqoop.

O Azure inclui muitos serviços que podem ser usados em uma arquitetura de Big Data. Eles se enquadram em aproximadamente duas categorias:

Serviços gerenciados, incluindo o Azure Data Lake Store, Azure Data Lake Analytics, Azure Synapse Analytics, Azure Stream Analytics, Hub de Eventos do Azure, Hub IoT do Azure e Azure Data Factory.

Tecnologias de software livre baseadas na plataforma Apache Hadoop, incluindo HDFS, HBase, Hive, Pig, Spark, Storm, Oozie, Sqoop e Kafka. Essas tecnologias estão disponíveis no Azure no serviço Azure HDInsight.

Essas opções não se excluem mutuamente e muitas soluções combinam tecnologias de software livre com serviços do Azure.

Quando usar essa arquitetura

Considere este estilo de arquitetura quando você precisar:

Armazenar e processar dados em volumes muito grandes para um banco de dados tradicional.

Transformar dados não estruturados para análise e relatório.

Capturar, processar e analisar fluxos não associados de dados em tempo real ou com baixa latência.

Usar Azure Machine Learning ou Serviços Cognitivos da Microsoft.

Benefícios

**Opções de tecnologia**. Você pode combinar gerenciados serviços do Azure e tecnologias Apache em clusters HDInsight para aproveitar recursos ou investimentos em tecnologia existentes.

**Desempenho por meio de paralelismo**. Soluções de Big Data aproveitam paralelismo, possibilitando soluções de alto desempenho dimensionadas para grandes volumes de dados.

**Escala elástica**. Todos os componentes da arquitetura de Big Data dão suporte a provisionamento de expansão para que você possa ajustar sua solução para cargas de trabalho grandes ou pequenas e pagar somente pelos recursos que usa.

**Interoperabilidade com soluções existentes**. Os componentes da arquitetura de Big Data também são usados para processamento IoT e soluções de BI empresariais, permitindo que você crie uma solução integrada entre cargas de trabalho de dados.

Desafios

**Complexidade**. Soluções de Big Data podem ser extremamente complexas, com vários componentes para lidar com a ingestão de dados de várias fontes de dados. Pode ser um desafio criar, testar e solucionar problemas de processos de Big Data. Além disso, pode haver um grande número de definições de configuração em vários sistemas que devem ser usados para otimizar o desempenho.

**Conjunto de qualificações**. Muitas tecnologias de Big Data são altamente especializadas e usam frameworks e idiomas que não são típicos de arquiteturas de aplicativo mais gerais. Por outro lado, as tecnologias de Big Data estão gerando novas APIs que se baseiam em linguagens mais estabelecidas. Por exemplo, a linguagem U-SQL no Azure Data Lake Analytics baseia-se em uma combinação de Transact-SQL e C#. Da mesma forma, APIs com base em SQL estão disponíveis para Hive, HBase e Spark.

**Maturidade da tecnologia**. Muitas das tecnologias usadas em Big Data estão em evolução. Embora tecnologias Hadoop centrais, como Hive e Pig, tenham se estabilizado, tecnologias emergentes, como Spark, apresentam grandes alterações e aprimoramentos a cada nova versão. Serviços gerenciados, como Azure Data Lake Analytics e Azure Data Factory, são relativamente jovens em comparação a outros serviços do Azure e provavelmente evoluirão ao longo do tempo.

**Segurança**. Soluções de Big Data normalmente se baseiam em armazenar todos os dados estáticos em um data lake centralizado. Proteger o acesso a esses dados pode ser desafiador, especialmente quando os dados devem ser ingeridos e consumidos por vários aplicativos e plataformas.

Práticas recomendadas

**Aproveitar o paralelismo**. A maioria das tecnologias de processamento de Big Data distribui a carga de trabalho em várias unidades de processamento. Isso exige que os arquivos de dados estáticos sejam criados e armazenados em um formato divisível. Sistemas de arquivos distribuídos, como HDFS, podem otimizar o desempenho de leitura e gravação, e o processamento real é executado por vários nós de cluster em paralelo, o que reduz o tempo de trabalho geral.

**Dados de partição**. O processamento em lote geralmente ocorre em um agendamento recorrente – por exemplo, semanal ou mensal. Arquivos de dados de partição e estruturas de dados como tabelas, com base em períodos de temporais que correspondem à agenda de processamento. Isso simplifica a ingestão de dados e o agendamento de trabalho, além de tornar mais fácil solucionar problemas de falhas. Além disso, o particionamento de tabelas usadas em consultas Hive, U-SQL ou SQL pode melhorar significativamente o desempenho da consulta.

**Aplicar semântica de esquema na leitura**. Usar um data lake permite combinar o armazenamento de arquivos em vários formatos, sejam estruturados, semiestruturados ou não estruturados. Use semântica de *esquema na leitura*, que projeta um esquema nos dados quando os dados estão sendo processados, não quando estão armazenados. Isso integra flexibilidade à solução e evita gargalos durante a ingestão de dados causados pela verificação de tipo e a validação de dados.

**Processar dados no local**. Soluções de BI tradicionais geralmente usam um processo ETL (extração, transformação e carregamento) para mover dados para um data warehouse. Com maiores volumes de dados e uma maior variedade de formatos, soluções de Big Data geralmente usam variações de ETL, como TEL (transformação, extração e carregamento). Com essa abordagem, os dados são processados no armazenamento de dados distribuídos, transformando-os na estrutura necessária, antes de mover os dados transformados para um armazenamento de dados analíticos.

**Equilibrar custos de tempo e utilização**. Para trabalhos de processamento em lotes, é importante considerar dois fatores: custo unitário de nós de computação e custo por minuto de usar esses nós para concluir o trabalho. Por exemplo, um trabalho em lotes pode levar oito horas com quatro nós de cluster. No entanto, pode ser que o trabalho use todos os quatro nós somente durante as primeiras duas horas, sendo apenas dois nós necessários depois disso. Nesse caso, executar todo o trabalho em dois nós aumentaria o tempo total do trabalho, mas não o duplicaria, de modo que o custo total seria menor. Em alguns cenários de negócios, um tempo de processamento mais longo pode ser preferível ao custo mais alto do uso de recursos de cluster subutilizados.

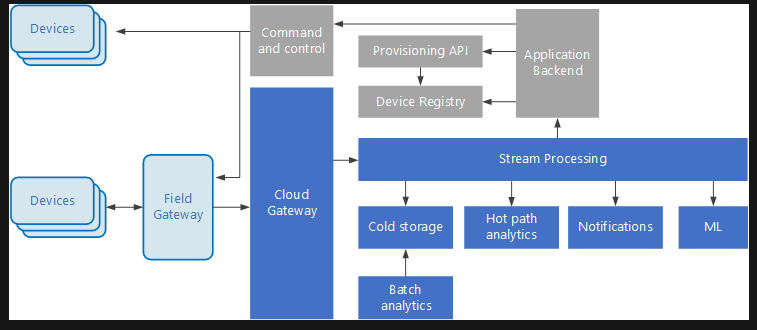
**Separar os recursos de cluster**. Ao implantar clusters HDInsight, você normalmente alcança um melhor desempenho provisionando recursos de cluster separados para cada tipo de carga de trabalho. Por exemplo, embora clusters do Spark incluam Hive, se você precisar executar amplo processamento com Hive e Spark, deverá considerar implantar clusters Spark e Hadoop dedicados separados. Da mesma forma, se você estiver usando HBase e Storm para processamento de fluxo de baixa latência e Hive para processamento em lotes, considere clusters separados para Storm, HBase e Hadoop.

**Orquestrar a ingestão de dados**. Em alguns casos, aplicativos de negócios existentes podem gravar arquivos de dados para processamento em lote diretamente em contêineres do Azure Storage Blob, em que podem ser consumidos pelo HDInsight ou pelo Azure Data Lake Analytics. No entanto, você geralmente precisará orquestrar a ingestão de dados de fontes de dados externas ou locais para o data lake. Use um fluxo de trabalho de orquestração ou um pipeline, como aqueles compatíveis com Azure Data Factory ou Oozie, para fazer isso de maneira previsível e gerenciável centralmente.

**Limpar dados confidenciais cedo**. O fluxo de trabalho de ingestão de dados deve remover dados confidenciais no início do processo para evitar armazená-los no data lake.

**Arquitetura do IoT**

O IoT (Internet das Coisas) é um subconjunto especializado de soluções de big data. O diagrama a seguir mostra uma possível arquitetura lógica de IoT. O diagrama enfatiza os componentes da arquitetura do streaming de eventos.



O gateway de nuvem consome eventos de dispositivo no limite da nuvem, usando um sistema de mensagens de latência baixa e confiável.

Os dispositivos podem enviar eventos diretamente para o gateway de nuvem, ou por meio de um gateway de campo. Um gateway de campo é um software ou dispositivo especializado, geralmente colocado com dispositivos, que recebe eventos e os encaminha para o gateway de nuvem. O gateway de campo também pode pré-processar os eventos de dispositivo brutos executando funções, como filtragem, agregação ou transformação de protocolo.

Após a ingestão, os eventos passam por um ou mais processadores de fluxo que podem encaminhar os dados (por exemplo, para armazenamento) ou executar análise e outros tipos de processamento.

A seguir estão alguns tipos comuns de processamento. (Esta lista certamente não é exaustiva.)

Gravando os dados de evento para armazenamento menos acessado, para arquivamento ou análise de processo em lote.

Análise de caminho mais acessado, analisando o fluxo de eventos (quase) em tempo real, para detectar anomalias, reconhecer padrões em janelas de tempo ou disparar alertas quando ocorre uma condição específica no fluxo.

Tratamento de tipos especiais de mensagens que não são de telemetria de dispositivos, como notificações e alarmes.

Machine Learning.

As caixas destacadas em cinza mostram os componentes de um sistema de IoT que não estão diretamente relacionadas ao streaming de evento, mas são incluídos aqui para fins de integridade.

O registro do dispositivo é um banco de dados dos dispositivos provisionados, incluindo os IDs de dispositivo e metadados do dispositivo, como localização.

A API de provisionamento é uma interface externa comum para provisionar e registrar dispositivos novos.

Algumas soluções IoT permitem que mensagens de comando e controle sejam enviadas aos dispositivos.

<https://learn.microsoft.com/pt-br/azure/architecture/guide/architecture-styles/big-data>

<https://rockcontent.com/br/blog/big-data/>