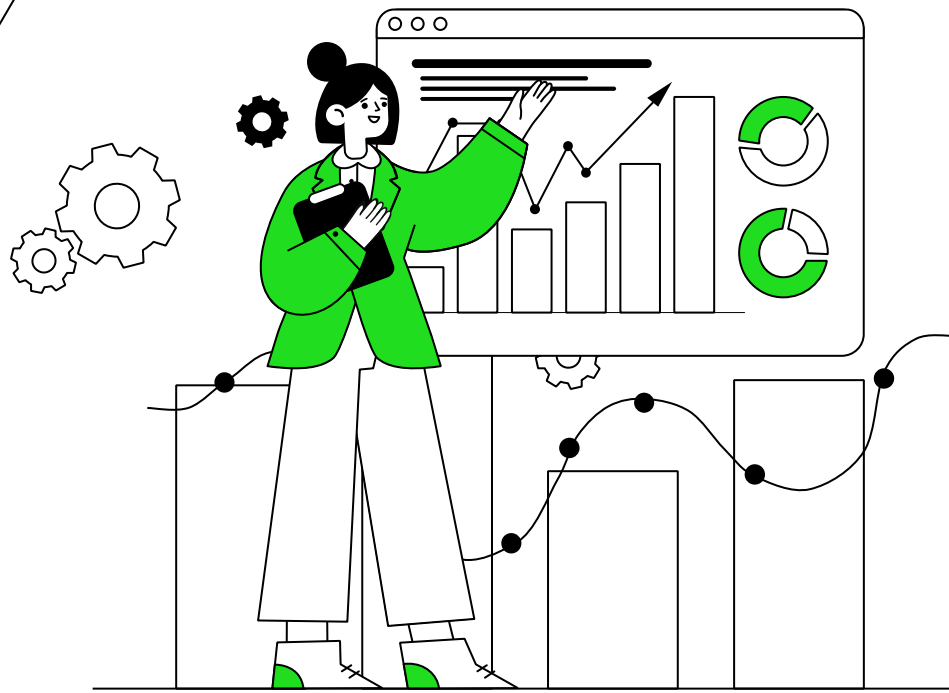
ARQUITETURA DE BIG DATA



O QUE É BIG DATA?

- Dados variados com volumes gradiosos e cresecentes que possuem velocidades crescentes.
- Softwares geralmente são incapazes de processar estes dados de grande porte.

BIG DATA ANALYTICS

São técnicas para analisar novos Insights e métricas, como aplicações que geram milhões de informações geradas a cada hora gerando valor de negócio. Esses dados devem ser agrupados, consolidados e analisados.

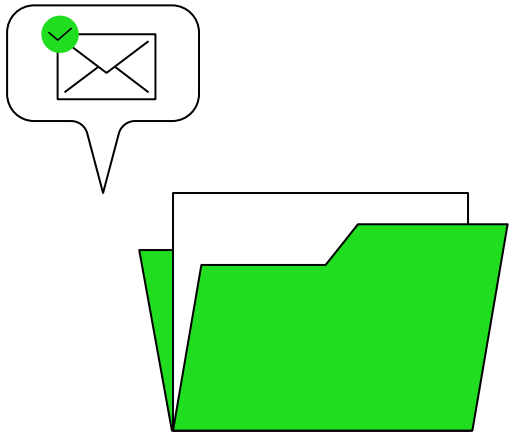
As técnicas de Big Data Analytics são:

- **Business Units:** avaliam as necessidades das àreas de négocio.
- **Data Science:** Uso de algoritmo para analisar, agrupar ou categorizar os dados.

OS TRÊS Vs DO BIG GATA

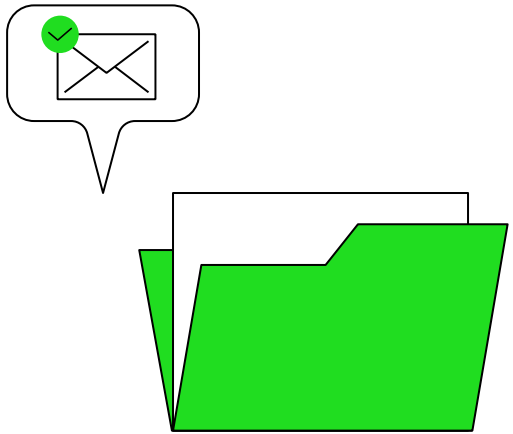
Os volumes do Big Data podem ser divididos em três, sendo eles:

- Velocidade
- Variedade
- Volume



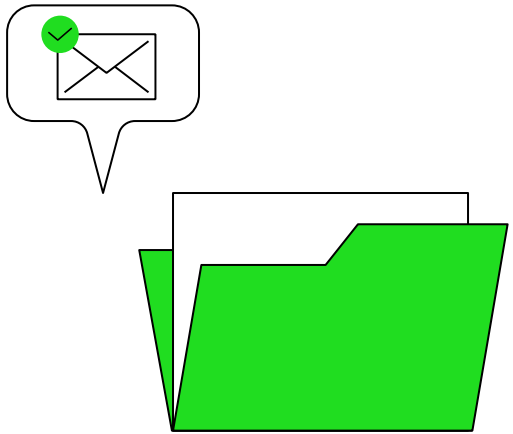
VELOCIDADE

- Taxa mais rápida pelo qual os dados são recebidos e possivelmente administrados.
- A velocidade dos dados é transmitida diretamente para a mémoria.
- Alguns produtos inteligentes são operados em tempo real ou quase em tempo real, exigindo avaliação e atuação em tempo real.



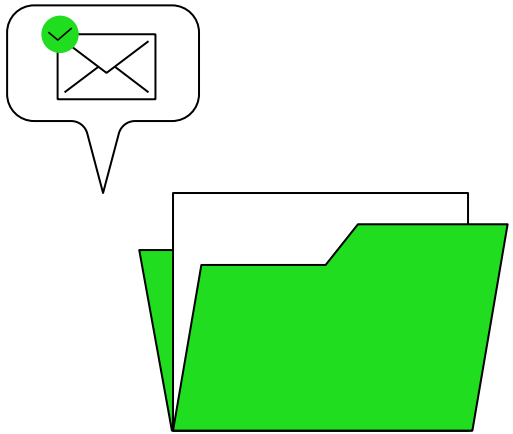
VOLUME

- A quantidade de dados importa
- A velocidade dos dados é transmitida diretamente para a mémoria.
- Alguns produtos inteligentes são operados em tempo real ou quase em tempo real, exigindo avaliação e atuação em tempo real.



VARIEDADE

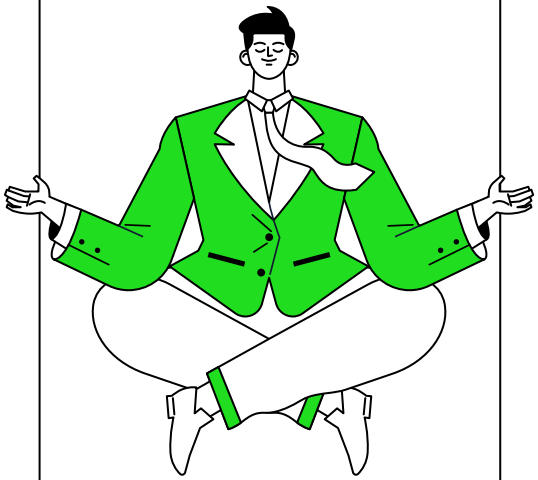
- Vários tipos de dados disponíveis
- Os dados podem ser estruturados, não estruturados e semiestruturados.
- Alguns produtos inteligentes são operados em tempo real ou quase em tempo real, exigindo avaliação e atuação em tempo real.



ESTRUTURADOS E NÃO ESTRUTURADOS

ESTRUTURADOS

Possuem algum padrão ou formato que pode ser usado na sua leitura e extração dos dados. Dados de bancos de dados, sistemas legados, arquivos texto (sejam csv, txt ou XML).



NÃO ESTRUTURADOS

Não possuem um formato padronizado para leitura, podem ser arquivos Word, Páginas de Internet/Intranet, Vídeos, áudios, entre outros

CASOS DE USO DO BIG DATA

01

DESENVOLVIMENTO DE PRODUTOS

Antecipar a demanda de clients criando modelos preditivos para novos produtos e serviços.

02

MANUTENÇÃO PREDITIVA

Análise de falhas mecânicas, investigando os possíveis problemas antes que eles ocorram.

03

EXPERIÊNCIA DO CLIENTE

Permite reunir dados dos clientes em mídias sociais, visitas a web, registro de chamadas e outras fontes para aprimorar a experiência.

04

FRAUDE E CONFORMIDADE

Identifica padrões em dados que podem resultar em fraudes.

05

MACHINE LEARNING

Ensinar máquinas a processar os dados.

06

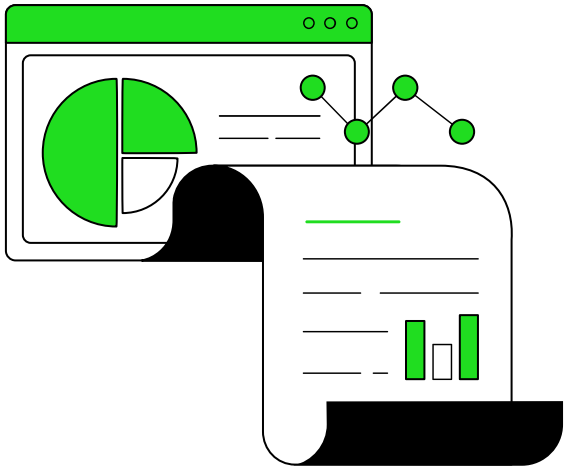
EFICIÊNCIA OPERACIONAL

Analisar e avaliar retornos dos clientes, produção e feedbacks com o intuit de reduzir e antecipar demandas futuras para aprimorar a tomada de decisão.

07

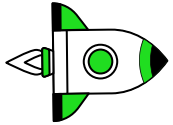
PROMOVA A INOVAÇÃO

Interdependências entre seres humanos, instituições, entidades e processos e, em seguida, determinando novas maneiras de usar esses insights. Usar informações de dados para aprimorar as decisões sobre considerações financeiras e de planejamento.



DESAFIOS DO BIG DATA

- O volume de dados dobra a cada ano.
- Os dados dependem da curadoria.



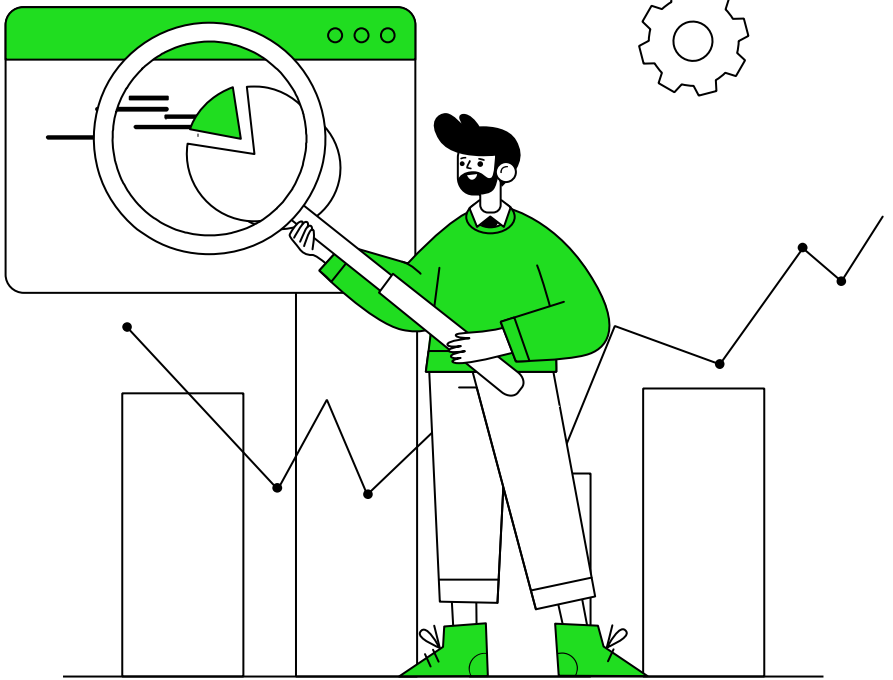
COMO FUNCIONA O BIG DATA?

Fornece novas informações que abrem novas oportunidades e modelos de negócio.

INTEGRAR

01

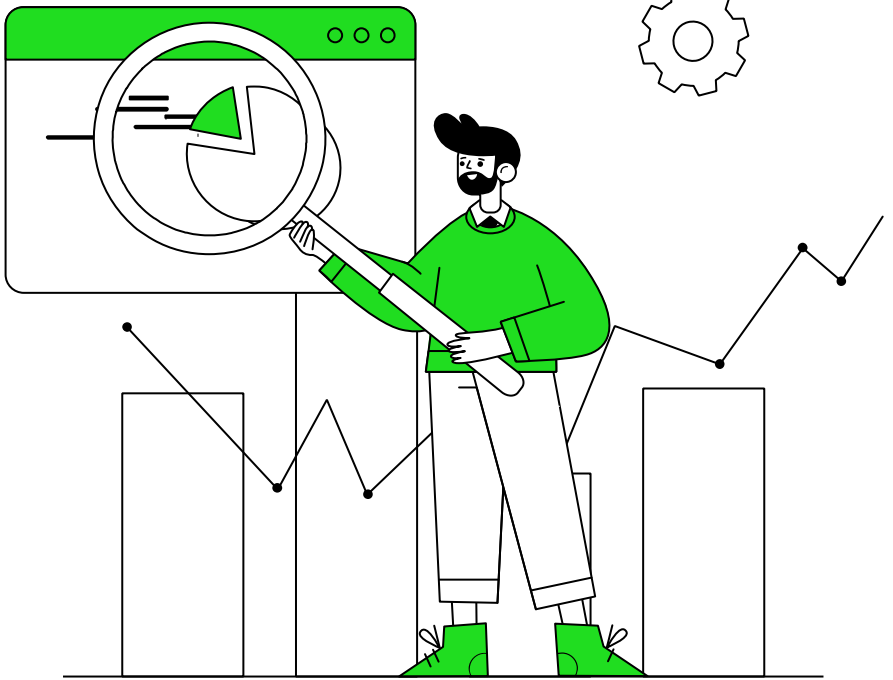
Durante a integração, você precisa inserir os dados, processá-los e verificar se estão formatados e disponíveis de forma que seus analistas de negócios possam começar a utilizá-lo.



GERENCIAR

02

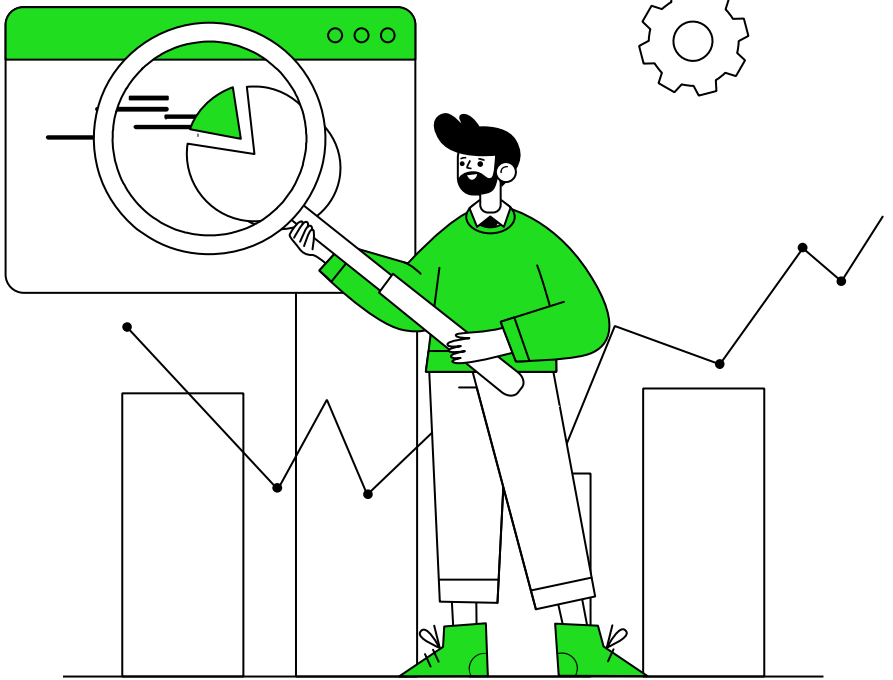
Big data exige armazenamento. Sua solução de armazenamento pode estar na nuvem, no local ou em ambos.



ANALISAR

03

Analisa os dados e obtém clareza, tendo um visual dos conjuntos de dados variados, explorando esses dados e tendo novas descobertas.



O QUE É ARQUITETURA DE BIG DATA?

Conjunto de modelos e regras que governam os dados e controlam como os dados coletados devem ser armazenados, organizados, integrados e usados nos sistemas de dados.

Uma arquitetura de dados deve definir padrões de dados para todos os sistemas de dados como uma visão ou um modelo das possíveis interações entre esses sistemas de dados.

A Arquitetura de Dados descreve como os dados são processados, armazenados
e utilizados em um sistema de informações.

ARQUITETURA
PARA BIG DATA

Projetada para lidar com ingestão, processamento e análise de dados grandes ou complexos demais para sistemas de banco de dados tradicionais.

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.

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.

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.

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.

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.

ANÁLISE E RELATÓRIO

A meta da maioria das soluções de Big Data é gerar insights sobre os dados por meio de análise e relatórios.

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.

PRÁTICAS RECOMENDADAS

APROVEITAR O PARALELISMO

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

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

PRÁTICAS RECOMENDADAS

APROVEITAR O PARALELISMO

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.

DADOS DE PARTIÇÃO

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

PRÁTICAS RECOMENDADAS

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.

DADOS DE PARTIÇÃO

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

PRÁTICAS RECOMENDADAS

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.

PRÁTICAS RECOMENDADAS

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

ORQUESTRAR A INGESTÃO DE DADOS

Em alguns casos, aplicativos de negócios existentes podem gravar arquivos de dados para processamento em lote diretamente em contêineres, 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