**ARQUITETURAS PARA BIG DATA**

# O que é big data?

A definição de big data são dados com maior variedade que chegam em volumes crescentes e com velocidade cada vez maior. Isso também é conhecido como os três Vs.

Simplificando, big data é um conjunto de dados maior e mais complexo, especialmente de novas fontes de dados. Esses conjuntos de dados são tão volumosos que o software tradicional de processamento de dados simplesmente não consegue gerenciá-los. No entanto, esses grandes volumes de dados podem ser usados para resolver problemas de negócios que você não conseguiria resolver antes.

O volume de dados gerado atualmente é monstruoso, todos os dias bilhões de novas informações são geradas globalmente, pense em todos os Apps, Sistemas, TVs, Celulares, aparelhos com IoT (Internet of Things ou Internet das Coisas) que estão capturando, processando e armazenando novos dados. Cada clique que é dado em uma página ou aplicativo é automaticamente guardado para que possa ser Analisado.

## Big Data Analytics ou Big Data e Inteligência Analítica

São todos os meios, técnicas e ferramentas utilizadas para analisar, tirar novos Insights (Ideias) e métricas de toda essa nova realidade dados, pensando por exemplo em um Aplicativo, como Waze ou Instagram, milhões de novas informações são geradas a cada hora e para gerar valor de negócio (Business Insights) esses dados precisam de alguma maneira serem agrupados, consolidados e principalmente Analisados.

Aqui entram as ferramentas e técnicas de Big Data Analytics, onde podemos atuar avaliando as necessidades das áreas de negócios (Business Units) ou aplicando algoritimos de Ciência de Dados (Data Science) para categorizar, agrupar e analisar esses dados.

## Os três Vs do big data

**Volume**

A quantidade de dados importa. Com o big data, você terá que processar grandes volumes de dados não estruturados de baixa densidade. Podem ser dados de valor desconhecido, como feeds de dados do Twitter, fluxos de cliques em uma página web ou em um aplicativo para dispositivos móveis, ou ainda um equipamento habilitado para sensores. Para algumas empresas, isso pode utilizar dezenas de terabytes de dados. Para outras, podem ser centenas de petabytes.

**Velocidade**

Velocidade é a taxa mais rápida na qual os dados são recebidos e talvez administrados. Normalmente, a velocidade mais alta dos dados é transmitida diretamente para a memória, em vez de ser gravada no disco. Alguns produtos inteligentes habilitados para internet operam em tempo real ou quase em tempo real e exigem avaliação e ação em tempo real.

**Variedade**

Variedade refere-se aos vários tipos de dados disponíveis. Tipos de dados tradicionais foram estruturados e se adequam perfeitamente a um banco de dados relacional. Com o aumento de big data, os dados vêm em novos tipos de dados não estruturados. Tipos de dados não estruturados e semiestruturados, como texto, áudio e vídeo, exigem um pré-processamento adicional para obter significado e dar suporte a metadados.

O que são informações estruturas e não estruturadas?

Consideramos:

Informações estruturadas – aquelas que 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).

Informações não estruturadas – não possuem um formato padronizado para leitura, podem ser arquivos Word, Páginas de Internet/Intranet, Vídeos, áudios, entre outros.

## A história do big data

Embora o conceito de big data em si seja relativamente novo, as origens de grandes conjuntos de dados remontam às décadas de 1960 e 1970, quando o mundo dos dados estava apenas começando, com os primeiros data centers e o desenvolvimento do banco de dados relacional.

Por volta de 2005, as pessoas começaram a perceber a quantidade de usuários de dados gerados pelo Facebook, YouTube e outros serviços online. O Hadoop (uma estrutura de código aberto criada especificamente para armazenar e analisar grandes conjuntos de dados) foi desenvolvido no mesmo ano. O NoSQL também começou a ganhar popularidade durante esse período.

O desenvolvimento de estruturas de código aberto, como o Hadoop, (e, mais recentemente, o Spark) foi essencial para o crescimento do big data, porque elas tornaram o trabalho com big data mais fácil e seu armazenamento mais barato. Nos anos seguintes, o volume de big data disparou. Usuários ainda estão gerando grandes quantidades de dados, mas não são somente humanos que estão fazendo isso.

Com o advento da Internet das Coisas (IoT), mais objetos e dispositivos estão conectados à internet, reunindo dados sobre padrões de uso do cliente e desempenho do produto. O surgimento do machine learning produziu ainda mais dados.

Apesar da evolução do big data, sua utilidade ainda está no começo. A computação em nuvem expandiu ainda mais as possibilidades do big data. A nuvem oferece uma escalabilidade verdadeiramente elástica, na qual os desenvolvedores podem simplesmente criar clusters ad hoc para testar um subconjunto de dados. E bancos de dados de grafos estão se tornando cada vez mais importantes também, com sua capacidade de exibir grandes quantidades de dados de uma forma que torna a análise rápida e abrangente.

## Benefícios do big data

Com o big data, você obtém respostas mais completas, porque tem mais informações.

Respostas mais completas significam mais confiança nos dados, ou seja, uma abordagem completamente diferente para lidar com problemas.

## Casos de uso de big data

O big data pode ajudar você a lidar com diversas atividades de negócios, desde a experiência do cliente até a análise avançada. Aqui estão algumas.

**Desenvolvimento de produtos**

Empresas como Netflix e Procter & Gamble usam big data para antecipar a demanda dos clientes. Eles criam modelos preditivos para novos produtos e serviços, classificando os principais atributos de produtos ou serviços passados e atuais e modelando a relação entre esses atributos e o sucesso comercial das ofertas. Além disso, a P&G utiliza dados e análises de grupos de foco, mídias sociais, mercados de teste e lançamentos antecipados de lojas para planejar, produzir e lançar novos produtos.

**Manutenção preditiva**

Fatores que podem prever falhas mecânicas podem estar profundamente relacionados a dados estruturados, como o ano, a marca e o modelo do equipamento, bem como em dados não estruturados que abrangem milhões de entradas de log, dados de sensores, mensagens de erro e temperatura do motor. Ao analisar essas indicações de possíveis problemas antes que eles ocorram, as empresas podem implementar a manutenção de maneira mais econômica e maximizar o tempo de atividade de peças e equipamentos.

**Experiência do cliente**

A corrida para alcançar os clientes já começou. Uma visão mais clara da experiência do cliente nunca foi tão possível como agora. O big data permite que você reúna dados de mídias sociais, visitas da web, registros de chamadas e outras fontes para aprimorar a experiência de interação e maximizar o valor fornecido. Comece a oferecer ofertas personalizadas, reduza a rotatividade de clientes e lide com problemas proativamente.

**Fraude e conformidade**

Quando se trata de segurança, você não está lidando apenas com alguns hackers mal-intencionados, e sim com equipes de especialistas na área. Os cenários de segurança e requisitos de conformidade estão evoluindo constantemente. Big data ajuda você a identificar padrões em dados que indicam fraudes e agregar grandes volumes de informações para tornar os relatórios regulamentares muito mais rápidos.

**Machine learning**

O machine learning é um dos assuntos mais comentados do momento. E dados, especificamente big data, é um dos motivos. Agora, somos capazes de ensinar máquinas em vez de programá-las. A disponibilidade de big data para treinar modelos de machine learning permite que isso seja uma realidade.

**Eficiência operacional**

A eficiência operacional nem sempre é notícia, mas é uma área em que o big data está tendo o maior impacto. Com big data, você pode analisar e avaliar a produção, feedbacks e retornos de clientes e outros fatores para reduzir interrupções e antecipar demandas futuras. Big data também pode ser usado para melhorar a tomada de decisões de acordo com a demanda atual do mercado.

**Promova a inovação**

O big data pode ajudar você a inovar, estudando interdependências entre seres humanos, instituições, entidades e processos e, em seguida, determinando novas maneiras de usar esses insights. Use informações de dados para aprimorar as decisões sobre considerações financeiras e de planejamento. Examine as tendências e o que os clientes desejam para oferecer novos produtos e serviços. Implemente um sistema de preços dinâmico. Há infinitas possibilidades.

## Desafios do big data

Embora o big data seja uma grande promessa, ele também traz seus desafios.

Para começar, o big data é… grande. Apesar de novas tecnologias terem sido desenvolvidas para o armazenamento de dados, os volumes de dados estão dobrando em tamanho a cada dois anos. As empresas ainda se esforçam para acompanhar a evolução de seus dados e encontrar maneiras de armazená-los com eficiência.

Mas armazenar os dados não é o suficiente. Eles devem ser usados para serem úteis, e isso depende da curadoria. Dados limpos ou relevantes para o cliente e organizados de maneira que permita uma análise significativa exigem muito trabalho. Cientistas de dados gastam de 50 a 80 por cento de seu tempo curando e preparando dados antes de serem usados.

Por fim, a tecnologia de big data está mudando em ritmo acelerado. Há alguns anos, o Apache Hadoop era a tecnologia popular usada para lidar com big data. Em seguida, o Apache Spark foi introduzido em 2014. Hoje, uma combinação das duas estruturas parece ser a melhor abordagem. Manter-se atualizado com a tecnologia de big data é um desafio contínuo.

## Como o big data funciona

O big data fornece novas informações que abrem novas oportunidades e modelos de negócios. Os primeiros passos envolvem três ações principais:

**1. Integrar**

O big data reúne dados de diversas fontes e aplicativos diferentes. Mecanismos tradicionais de integração de dados, como extrair, transformar e carregar (ETL), geralmente não estão aptos à tarefa. Isso requer novas estratégias e tecnologias para analisar conjuntos de big data em terabytes ou até mesmo em escala de petabytes.

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-los.

**2. Gerenciar**

Big data exige armazenamento. Sua solução de armazenamento pode estar na nuvem, no local ou em ambos. Você pode armazenar seus dados da forma que desejar e trazer os requisitos de processamento desejados e os mecanismos de processo necessários para esses conjuntos de dados sob demanda. Muitas pessoas escolhem a solução de armazenamento de acordo com a localização atual dos dados. A nuvem está gradualmente ganhando popularidade porque é compatível com as suas necessidades atuais de computação e permite que você crie recursos conforme necessário.

**3. Analisar**

Seu investimento em big data é compensado quando você analisa seus dados e age com base neles. Obtenha mais clareza com uma análise visual dos seus conjuntos de dados variados. Explore ainda mais os dados para fazer novas descobertas. Compartilhe suas descobertas com os outros. Crie modelos de dados com machine learning e inteligência artificial. Faça seus dados funcionarem.

Ref.:

[**https://www.oracle.com/br/big-data/what-is-big-data/**](https://www.oracle.com/br/big-data/what-is-big-data/)

<https://cetax.com.br/big-data/>

<http://blog.marcoreis.net/arquitetura-de-referencia-para-solucoes-de-big-data/>

# O que é Arquitetura de Dados?

Arquitetura de Dados é um 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 de uma organização. Em suma, a Arquitetura de Dados define padrões para todos os sistemas de dados como uma visão ou um modelo do funcionamento das interações dos 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. Uma arquitetura de dados, em parte, descreve as estruturas de dados usadas por uma empresa. As arquiteturas de dados abordam dados em armazenamento, dados em uso e dados em movimento; descrições de armazenamentos de dados, grupos de dados e itens de dados; e mapeamentos desses artefatos de dados para controle de qualidade, gestão de metadados, etc.

A Arquitetura de Dados descreve como os dados são processados, armazenados e utilizados em um sistema de informações. Ele fornece critérios para operações de processamento de dados, a fim de possibilitar o design de fluxos de dados e também controlar o fluxo de dados no sistema.

E à medida a Ciência de Dados cresce, os desafios associados em construir uma arquitetura que permita extrair o melhor dos dados, cresce na mesma proporção.

– Analisar e modelar os dados é fundamental (papel do Cientista de Dados).

– Executar os pipelines de dados é fundamental (papel do Engenheiro de Dados).

– Construir um fluxo de aplicação de modelos preditivos é fundamental (papel do Engenheiro de Machine Learning).

– Desenvolver aplicações baseadas em Inteligência Artificial é fundamental (papel do Engenheiro de IA).

– Automatizar todo esse processo é fundamental (papel do Engenheiro Data Ops).

Mas quem vai olhar para tudo isso, planejando a integração com todos os demais sistemas da empresa e garantindo que os regulamentos de proteção aos dados estejam sendo seguidos? Entra em cena o Arquiteto de Dados.

Ref.:

<https://blog.dsacademy.com.br/o-que-e-arquitetura-de-dados/>

# Arquitetura para big data

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. O limite no qual as organizações ingressam no campo do Big Data é diferente, dependendo das capacidades dos usuários e de suas ferramentas. Para alguns, isso pode significar centenas de gigabytes de dados, enquanto para outros, centenas de terabytes. À medida que as ferramentas para o trabalho com conjuntos de Big Data evoluem, na mesma proporção evolui o significado de Big Data. Cada vez mais, esse termo se relaciona ao valor que é possível extrair dos conjuntos de dados por meio de análise avançada, em vez de estritamente o tamanho dos dados, embora nesses casos, eles tendam a ser muito grandes.

Ao longo dos anos, o cenário dos dados vem mudando. Houve uma mudança no que você pode fazer ou o que deve fazer, com os dados. O custo de armazenamento caiu drasticamente, enquanto os meios pelos quais os dados são coletados continuam aumentando. Alguns dados são recebidos a um ritmo rápido, constantemente exigindo sua coleta e observação. Outros dados são recebidos mais lentamente, mas em partes muito grandes, geralmente na forma de décadas de dados históricos. Talvez você esteja enfrentando um problema de análise avançada ou um problema que exija o aprendizado de máquina. Esses são desafios que as arquiteturas de Big Data buscam resolver.

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.

## 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.

Ref.:

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

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