**O que é Big Data Architecture?**

A arquitetura Big Data é o sistema global usado para gerenciar grandes quantidades de dados para que possam ser analisados para fins comerciais, conduzir a análise de dados e fornecer um ambiente no qual as grandes ferramentas de análise de dados podem extrair informações comerciais vitais de dados ambíguos. Serve como um plano de referência para a infraestrutura de dados e soluções, a fim de que definam logicamente como grandes soluções de dados irão funcionar e os detalhes de segurança.

A arquitetura de Big Data possui quatro camadas lógicas, sendo elas:

**Big Data Sources Layer:** ambiente de big Data pode gerenciar tanto o processamento em lote quanto o processamento em tempo real de grandes fontes de dados.

**Management & Storage Layer(Camada de Armazenamento e Administração):** recebe dados da fonte, converte os dados em um formato compreensível para a ferramenta de análise de dados, e armazena os dados de acordo com seu formato.

**Analysis Layer(Camada de Análise):** as ferramentas analíticas extraem a inteligência comercial da grande camada de armazenamento de dados.

**Consumption Layer (Camada de Consumo):** recebe resultados da grande camada de análise de dados e os apresenta à camada de saída pertinente - também conhecida como camada de inteligência empresarial.

**Processos Big Data Arquitetura**

**Fontes de Dados:** Todas as fontes que alimentam a extração de dados estão sujeitas a esta definição, portanto é aqui que se encontra o ponto de partida para big Data. Como por exemplo banco de dados, data warehouses, data warehouses baseados em nuvem, aplicações SaaS, dados em tempo real de servidores e sensores, como dispositivos IoT

**Armazenamento de dados:** É possível armazenar muitos arquivos grandes baseados em formatos diferentes no data lake. Isto consiste nos dados que são gerenciados para operações em lote e que são salvos nas lojas de arquivos. Fornecemos armazenamento de HDFS, Microsoft Azure, AWS, e GCP, entre outros.

**Processamento em lote:** Cada pedaço de dados é dividido em diferentes categorias usando long-running jobs, que filtram, agregam e preparam os dados para análise. Esses jobs normalmente requerem fontes, processa-os e entrega-os a novos arquivos.

**Ingestão de mensagens com base em tempo real:** O processo de entrega, juntamente com outras semânticas de enfileiramento de mensagens, é geralmente mais confiável.

O processamento de mensagens em tempo real utiliza os dados ingeridos como uma ferramenta de publicação e assinatura, enquanto leva em conta todos os dados ingeridos em primeiro lugar e depois os utiliza como uma ferramenta de publicação e assinatura. O processamento do fluxo, por outro lado, trata todos esses dados de fluxo sob a forma de janelas ou fluxos e os grava na pia.

**Baseado em dados analíticos:** Para analisar e processar dados já processados, as ferramentas analíticas utilizam o armazenamento de dados baseado no HBase ou qualquer outra tecnologia de armazenamento de dados NoSQL.

**Orquestração:** Soluções baseadas em dados que utilizam grandes dados são tarefas relacionadas a dados que são repetitivos por natureza, e que também estão contidas em cadeias de fluxo de trabalho que podem transformar os dados da fonte e também mover dados através de fontes, bem como pias e cargas em lojas.

**Tipos de Arquitetura de Big Data**

**Arquitetura Lambda**

Uma única arquitetura Lambda lida tanto com dados de lote (estáticos) quanto com dados de processamento em tempo real. Ela é empregada para resolver o problema da computação de funções arbitrárias.

**Arquitetura Kappa**

Substitui o meio de fornecimento de dados por filas de mensagens. Os mecanismos de mensagens armazenam uma sequência de dados nos bancos de dados analíticos, que são então lidos e convertidos no formato apropriado antes de serem salvos para o usuário final.

A arquitetura facilita o acesso à informação em tempo real, lendo e transformando os dados das mensagens em um formato facilmente acessível aos usuários finais.

**Ferramentas e Técnicas de Big Data**

**Processamento Massivamente Paralelo (MPP)**

Um sistema de armazenamento de nada acoplado ou compartilhado é uma construção de processamento maciçamente paralelo com o objetivo de dividir um grande número de máquinas de computação em peças discretas e proceder em paralelo. Um sistema MPP também é referido como um sistema de nada acoplado frouxamente ou nada compartilhado. O processamento é realizado dividindo um grande número de processadores de computador em partes separadas e procedendo em paralelo.

**Bases de dados não-SQL**

Estruturas são empregadas para ajudar a associar dados com um determinado domínio. Os dados não podem ser armazenados em um banco de dados estruturado, a menos que sejam primeiro convertidos em um. SQL (ou NoSQL) é uma linguagem não estruturada utilizada para encapsular dados não estruturados e criar estruturas para dados heterogêneos no mesmo domínio.

**Aplicação de Arquitetura de Big Data**

A estrutura de uma arquitetura de Big Data permite a exclusão de dados sensíveis logo no início por causa de seu procedimento de ingestão de dados e por causa de seu armazenamento em Data Lake.O processamento em lote tem uma frequência e um cronograma recorrente. O processo de ingestão e o cronograma de trabalho para os dados do lote são simplificados, pois os arquivos de dados podem ser particionados. O desempenho da consulta é melhorado com o particionamento das tabelas. Colmeia, U-SQL ou consultas SQL são usadas para particionar os dados das tabelas.Os arquivos de lote distribuídos podem ser divididos ainda mais usando paralelismo e tempo de trabalho reduzido. Outra aplicação é dispersar a carga de trabalho entre as unidades de processamento. Os arquivos de lotes estáticos são criados e salvos em formatos que podem ser divididos ainda mais.

**Benefícios da Arquitetura de Big Data**

**Computação paralela de alto desempenho:** Arquitetura de Big Data empregam a computação paralela, na qual servidores multiprocessadores realizam muitos cálculos ao mesmo tempo para acelerar o processo.

**Escalabilidade elástica:** Arquitetura de Big Data podem ser escalonadas horizontalmente, permitindo que o ambiente seja ajustado ao tamanho das cargas de trabalho. Uma grande solução de dados é normalmente operada na nuvem.

**Liberdade de escolha:** Grandes arquiteturas de dados podem usar várias plataformas e soluções no mercado pode-se escolher a combinação certa de soluções para suas cargas de trabalho específicas, sistemas existentes e níveis de especialização em TI para alcançar o melhor resultado.

**A capacidade de interoperar com outros sistemas:** Usar componentes de arquitetura de Dados Grandes para processamento e BI, e também como fluxos de trabalho analíticos.

**Desafios de Big Data**

**Segurança:** É necessária uma segurança robusta. Além disso, o estabelecimento de um acesso seguro pode ser difícil. Outras aplicações também devem consumir dados para que eles funcionem.

**Complexidade:** As partes móveis de uma arquitetura de Big Data normalmente consistem de muitos elementos interligados. Estes componentes podem ter seus próprios dutos de injeção de dados e várias configurações para melhorar o desempenho, além de muitas intervenções de configuração de componentes cruzados.

**Tecnologias em evolução:** Escolher as soluções e componentes certos é fundamental para atender aos objetivos comerciais de Big Data.

**Expertise em um domínio específico:** À medida que as Big Data APIs são construídas sobre as principais linguagens se tornam gradualmente populares. As arquiteturas e soluções de Big Data geralmente usam linguagens e estruturas únicas e altamente especializadas que impõem uma curva de aprendizado substancial tanto para desenvolvedores quanto para analistas de dados.

Aplicações de Big Data

**Grandes Dados na Indústria Bancária e de Títulos**

O setor bancário utiliza Big Data para análise de risco, anti-lavagem de dinheiro e outras fraudes financeiras desse tipo, portanto, tais fraudes são rapidamente detectadas e mitigadas.

**Grandes Dados em Comunicações, Mídia e Entretenimento**

O conteúdo da mídia é conduzido pela análise dos dados dos clientes na escala da Big Data para compreender os padrões de comportamento, gostos e aversões de um amplo conjunto de clientes.

**Big Data na área da saúde**

Big Data é usado para fornecer diagnósticos baseados em evidências ao invés de uma bateria de testes médicos, trazendo assim custos e melhorando a eficiência nos círculos médicos.

**Grandes dados na fabricação**

Big Data é amplamente utilizado na fabricação para Gerenciamento da Cadeia de Suprimentos, Manutenção Preditiva, Qualidade Preditiva, Previsão de Produção, melhoria da produção e rendimento.

**Grandes Dados em Seguros**

Big Data ajuda as companhias de seguros a minimizar os riscos de subscrição e a melhorar a detecção de fraudes.

**Comércio Varejista**

A maior aplicação de Big Data é provavelmente na indústria de varejo e comércio eletrônico, com análise de dados de vários fluxos, incluindo as mídias sociais, para a produção de anúncios direcionados.

Mais Informações em:

<https://www.heavy.ai/technical-glossary/big-data-architecture>

<https://www.interviewbit.com/blog/big-data-architecture/>

https://www.jigsawacademy.com/blogs/big-data/big-data-architecture/