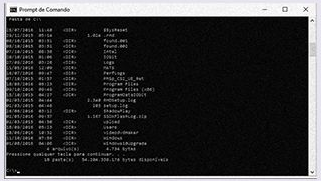
programa{m}aria
/sprint

26 de **setembro** a 04 de **outubro**

online

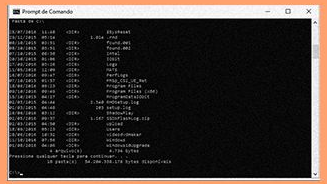


IA e Dados

programa{m}aria
/sprint

PALESTRA

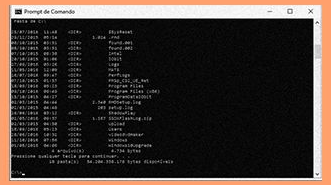
Arquitetando Dados

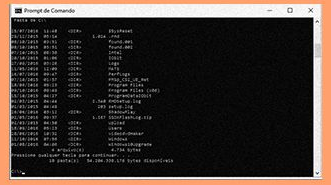


Débora Mariano

Data Engineer Lead – Localiza & CO

programa{m}aria
/sprint





Arquitetura de Dados: Definição e Beneficios

A Arquitetura de Dados é a estrutura e a organização dos dados dentro de uma empresa. Ela define como os dados são coletados, armazenados, gerenciados e utilizados. A arquitetura de dados é essencial para garantir que os dados sejam acessíveis, precisos e seguros, permitindo que as empresas tomem decisões informadas e estratégicas. Ela é fundamental para as operações de processamento de dados e aplicativos de inteligência artificial (IA).

A arquitetura de dados oferece diversos benefícios para as empresas, ajudando a gerenciar e utilizar informações de maneira eficaz. Aqui estão alguns dos principais benefícios:

#1 **Melhoria na Tomada de Decisões**: Acesso a dados precisos e atualizados, facilitando decisões baseadas em dados.

#2 **Eficiência Operacional**: Redução de redundâncias e inconsistências, otimizando processos e aumentando a eficiência operacional.

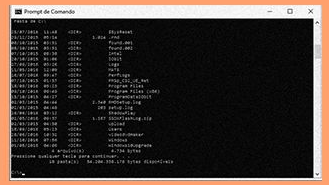
#3 **Segurança e Conformidade**: Garante a proteção das informações sensíveis e ajuda a cumprir regulamentos de privacidade e segurança de dados

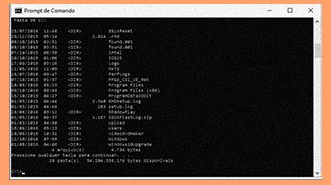
#4 **Redução de Custos**: Redução custos associados ao armazenamento e processamento de dados.

#5 **Visão Integrada**: Facilita a compreensão dos dados do negócio, permitindo uma visão única e integrada que apoia a estratégia empresarial

#6 **Suporte à Inovação**: Com dados bem gerenciados, as empresas podem explorar novas tecnologias e iniciativas, como inteligência artificial e análise avançada de dados

programa{m}aria
/sprint





Arquitetura de Dados: Principais componentes

Modelagem de Dados

Existem três tipos principais :

#1 **Conceitual**

#2 **Lógico**

#3 **Físico**

Modelos de dados são representações abstratas que descrevem a estrutura, as relações e as restrições dos dados em um sistema de informação.

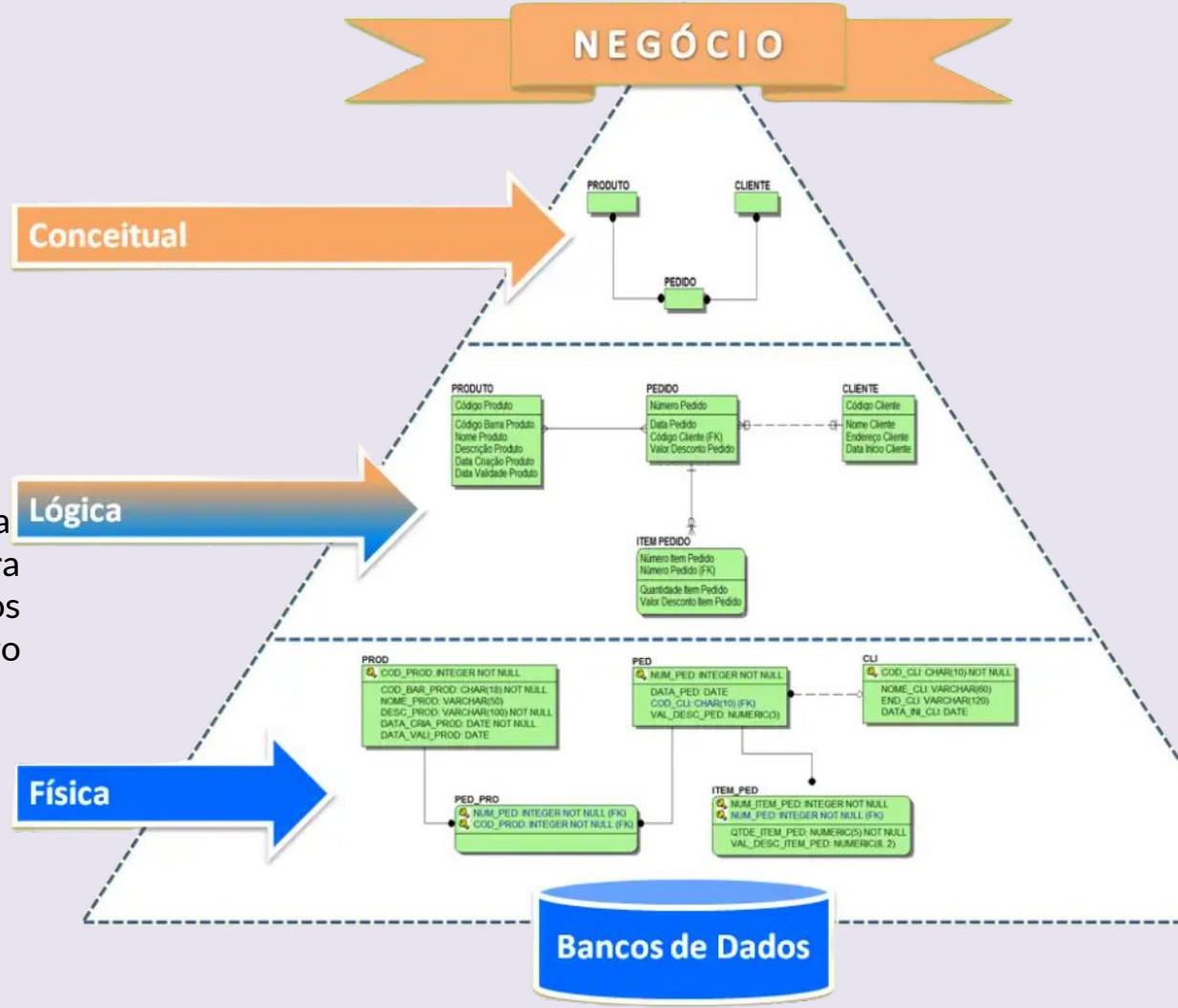
Eles são fundamentais para a organização e o gerenciamento eficaz dos dados.

Modelagem de Dados

O modelo de Dados conceitual é uma representação de alto nível dos dados de um sistema, focando nos conceitos e nas relações entre eles, sem se preocupar com detalhes técnicos ou de implementação.

O modelo de Dados Lógico é uma representação detalhada da estrutura dos dados, focando em como os dados são organizados e relacionados dentro de um sistema

O modelo de Dados físico representa os dados serão armazenados fisicamente no banco de dados. Isso inclui a definição de tabelas, índices, partições, e outros elementos de armazenamento.



Coleta de Dados

A coleta de dados é uma etapa crucial na arquitetura de dados, pois envolve a captura de informações de diversas fontes para posterior processamento e análise.

Identificação de Fontes de Dados

A primeira etapa na coleta de dados é identificar de onde os dados serão coletados. As fontes podem ser variadas, incluindo:

- **Sistemas Internos**: Bancos de dados transacionais, CRM entre outros.
- **Dispositivos**: Sensores e dispositivos conectados que coletam dados em tempo real.
- **APIs (Application Programming Interfaces)**: Integração e troca de dados entre diferentes sistemas e aplicações.
- **Mídias Sociais**: Dados de plataformas como Facebook, Twitter, LinkedIn.
- **Streaming de Dados**: Dados gerados e processados em tempo real, como logs de servidores e transações financeiras.
- **Arquivos e Documentos**: Dados armazenados em arquivos de texto, planilhas, PDFs.

Processamento de Dados

Vamos detalhar os principais métodos de coleta e processamento de dados:

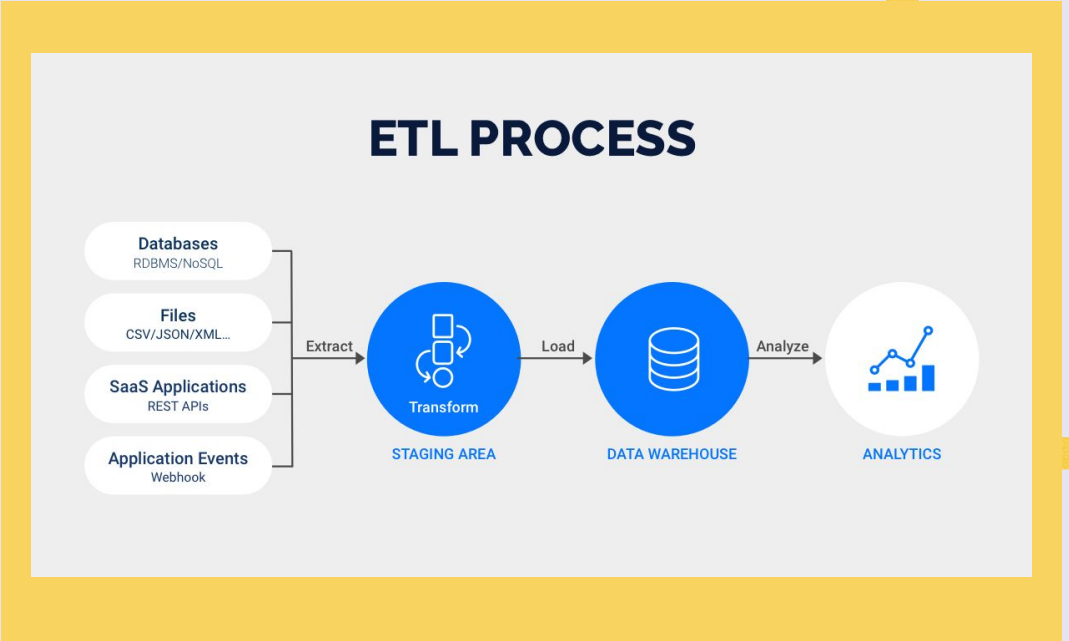
- **ETL (Extract, Transform, Load)**: Processo tradicional que envolve a extração de dados de várias fontes, transformação para um formato adequado e carregamento em um data warehouse ou data lake.

- **Streaming de Dados**: Coleta de dados em tempo real de fontes como sensores IoT, logs de servidores e feeds de redes sociais

- **APIs**: Utilização de APIs para integrar e coletar dados de diferentes sistemas e serviços externos*3*.

Extract,Transform and Load

ETL (*Extract*, *Transformation*, *Load*) que significa Extração, Transformação e Carga, é um processo utilizado para transferir dados de diferentes fontes, convertê-los em um formato adequado e carregá-los em um sistema de destino, como um *data warehouse* ou um banco de dados analítico. O objetivo final do ETL é fornecer dados consistentes, de alta qualidade e prontos para análise.



Armazenamento de Dados

O armazenamento de dados é uma parte crucial da arquitetura de dados, pois envolve a preservação e organização das informações para que possam ser acessadas e utilizadas de maneira eficiente. Vamos detalhar os principais aspectos do armazenamento de dados:

Armazenamento de Dados

O armazenamento de dados é uma parte crucial da arquitetura de dados, pois envolve a preservação e organização das informações para que possam ser acessadas e utilizadas de maneira eficiente. Vamos detalhar os principais tipos de componentes de armazenamento e de arquitetura dados :

Data Warehouse

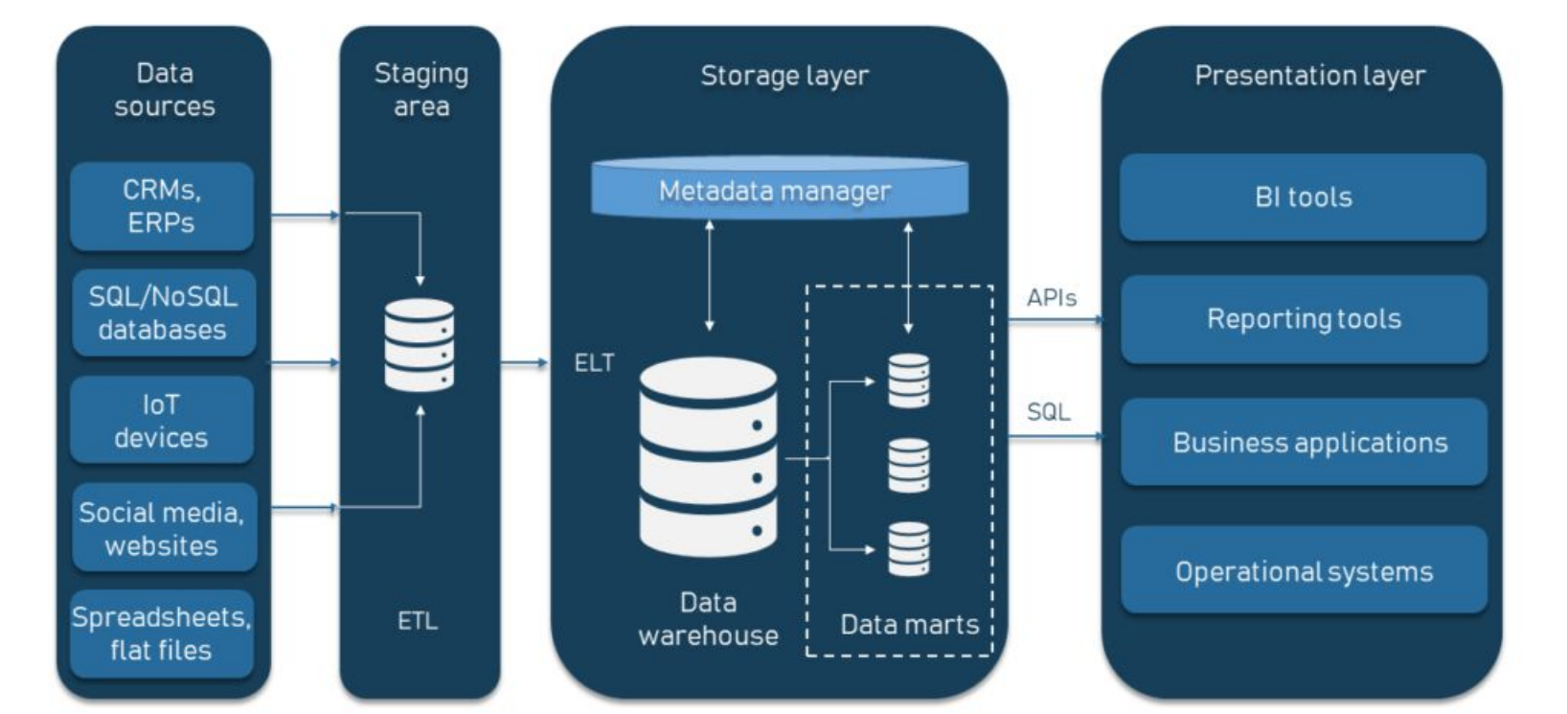
Agrega dados de diferentes fontes de dados relacionais de uma empresa em único repositório, consistente e centralizado. Após a extração, os dados fluem por um pipeline de dados ETL, passando por diversas transformações para atender ao modelo de dados predefinido. Uma vez carregados no data warehouse, os dados são encarregados de apoiar diferentes aplicativos de inteligência de negócios (BI) e ciência de dados.

Algumas ferramentas de mercado: Amazon Redshift, Google BigQuery

Data Marts

Um data mart é uma versão centrada um data warehouse que contém um subconjunto menor de dados importantes e necessários para uma única equipe ou um grupo específico de usuários dentro de uma organização. Por conter um subconjunto menor de dados, os data marts permitem que um departamento ou linha de negócios descubram insights especializados mais rapidamente em comparação com o conjunto mais amplo de dados de um data warehouse.

Data Warehouse + Data Marts



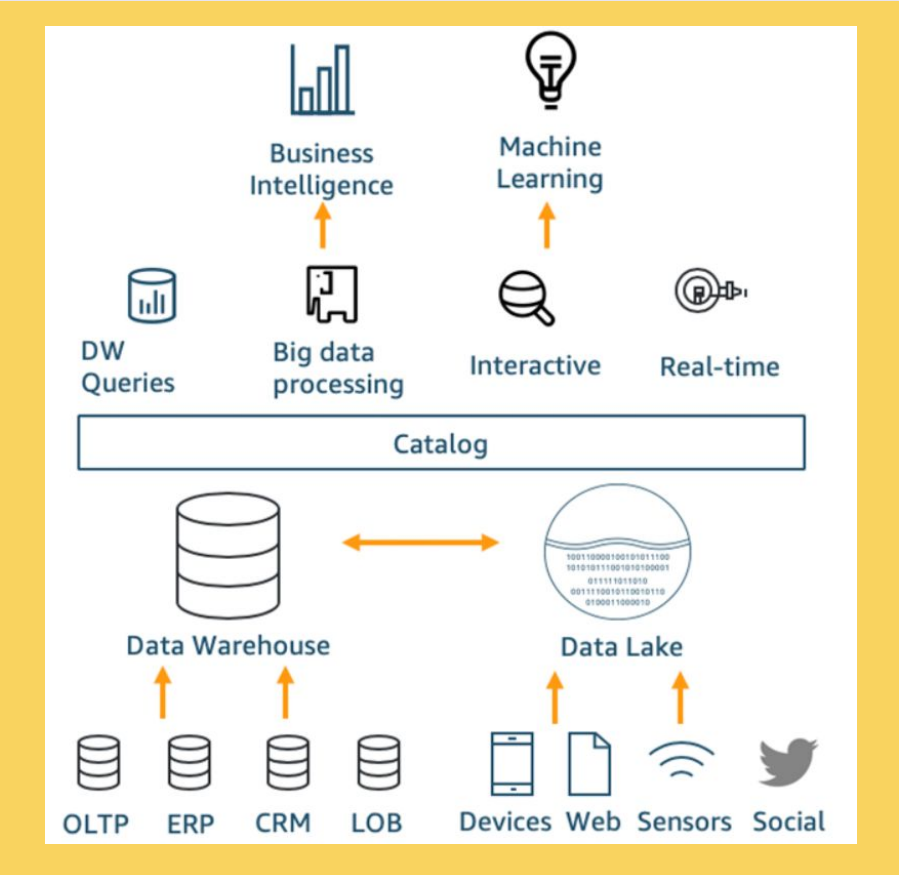
Data Lake

Um **data lake** é um repositório centralizado que permite armazenar grandes volumes de dados em seu formato bruto, sejam eles estruturados, semiestruturados ou não estruturados.

Algumas ferramentas de mercado: Amazon S3, Azure Data Lake Storage, Hadoop Distributed File System (HDFS).

Data Lake + Data Warehouse

Diferente dos data warehouses, que armazenam dados estruturados, os data lakes podem armazenar dados em seu formato original, sem a necessidade de transformação prévia.



Banco de Dados relacionais e não relacionais

Um **banco de dados relacional** é um conjunto de informações que organiza dados em relações predefinidas, em que os dados são armazenados em uma ou mais tabelas (ou "relações") de colunas e linhas.
Exemplos: MySQL, PostgreSQL, Oracle, SQL Server.

Um **banco de dados não relacional** ou **NoSQL**, ou seja, que não seguem o modelo de tabelas e relacionamentos utilizado pelos bancos de dados relacionais tradicionais.
Exemplos: MongoDB, Cassandra, Redis*3*.

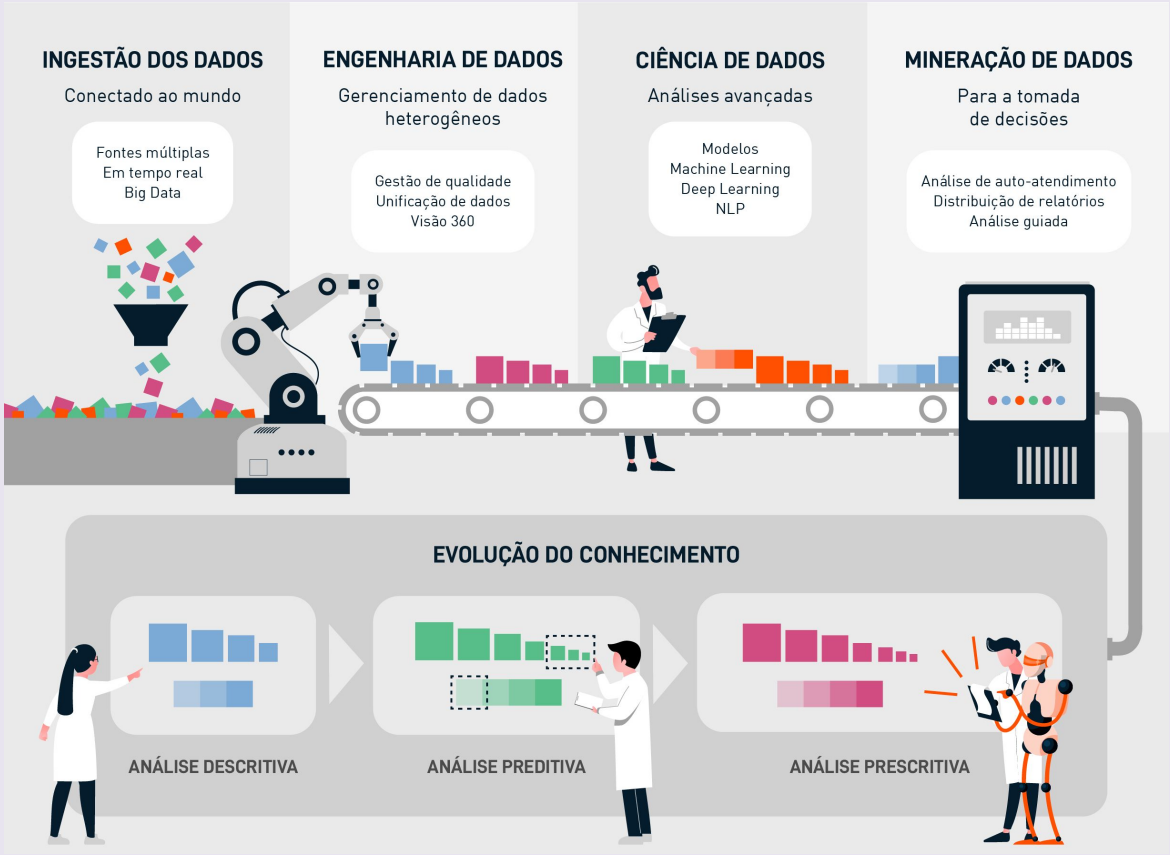
Banco de Dados relacionais x não relacionais

Quando usar?

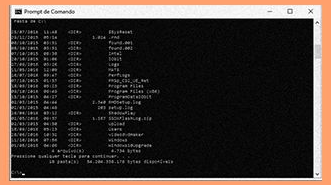
Bancos de dados relacionais são a melhor opção quando seus dados são previsíveis em termos de tamanho, estrutura e frequência de acesso ou se os relacionamentos entre entidades forem importantes. Por exemplo, se você tem um grande conjunto de dados com uma estrutura e relacionamentos complexos, quer que esses relacionamentos se destaquem pela análise e facilidade de uso.

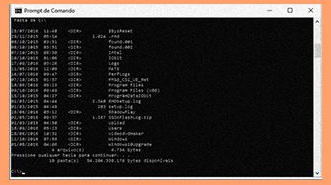
Por outro lado, um modelo não relacional funciona melhor para armazenar dados flexíveis em forma ou tamanho, ou que possam mudar no futuro.

Da coleta ao consumo dos dados ...



programa{m}aria
/sprint





Arquitetura de Dados: Caso de Uso - Netflix

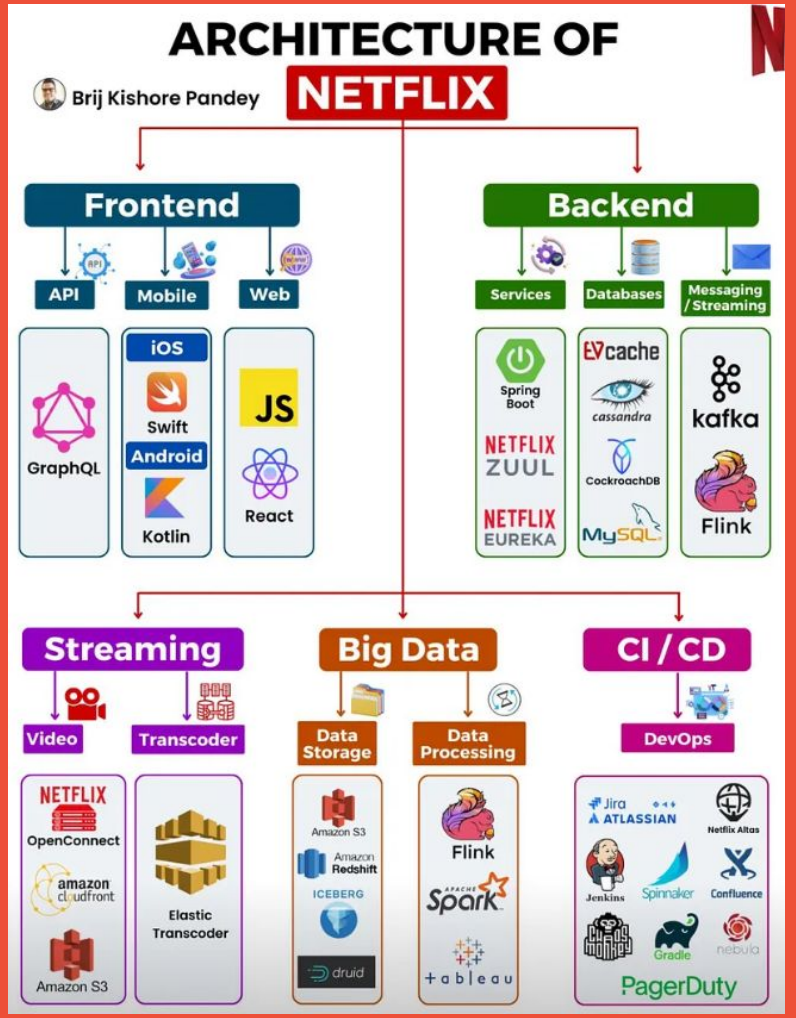
Caso de Uso: Netflix

A Netflix é uma das maiores plataformas de streaming do mundo, com milhões de usuários globais. A Netflix utiliza dados de maneira estratégica para se diferenciar no mercado de streaming.

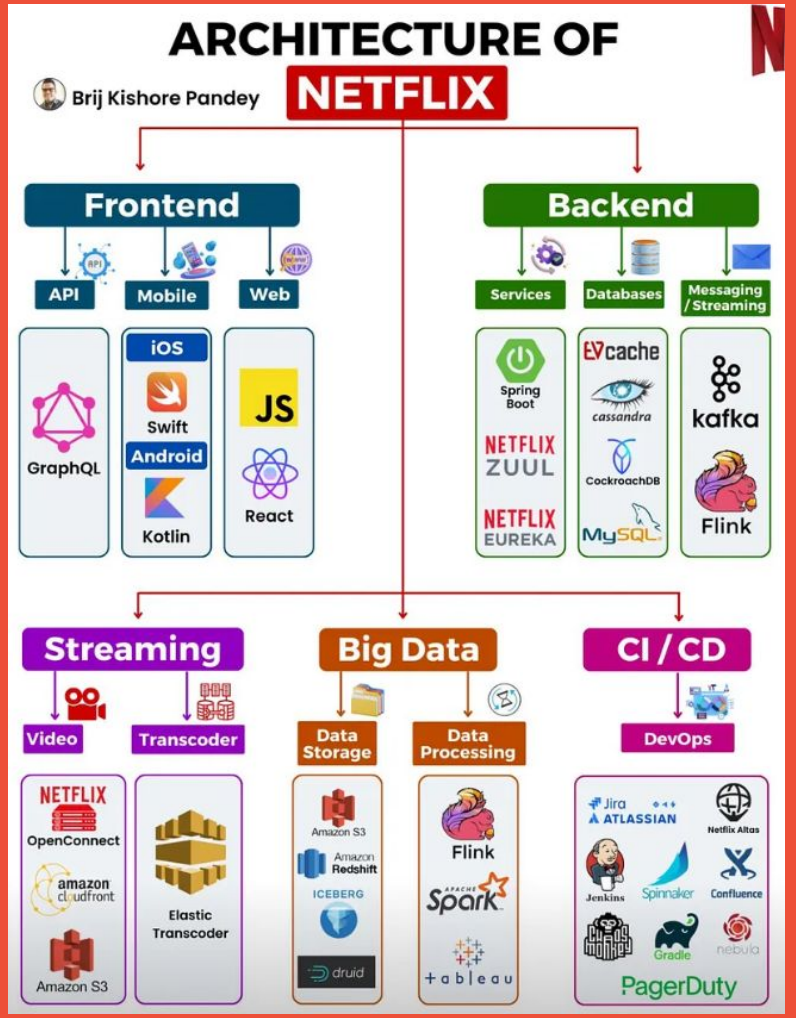
Desafio

Personalizar recomendações de conteúdo para cada usuário, melhorando a experiência do cliente e aumentando o tempo de visualização. A Netflix lida com grandes volumes de dados para personalizar recomendações de conteúdo e aprimorar experiências do usuário.

Banco de Dados: A infraestrutura de dados massiva da Netflix exige um sistema de gerenciamento de banco de dados robusto. Para cache, a Netflix usa o **EVCache**, um cache distribuído na memória, para otimizar o acesso aos dados e os tempos de resposta. Além disso, a Netflix aproveita a escalabilidade e a resiliência do **cockroachDB e do MySQL** para seus principais requisitos de banco de dados, fornecendo gerenciamento de dados perfeito.

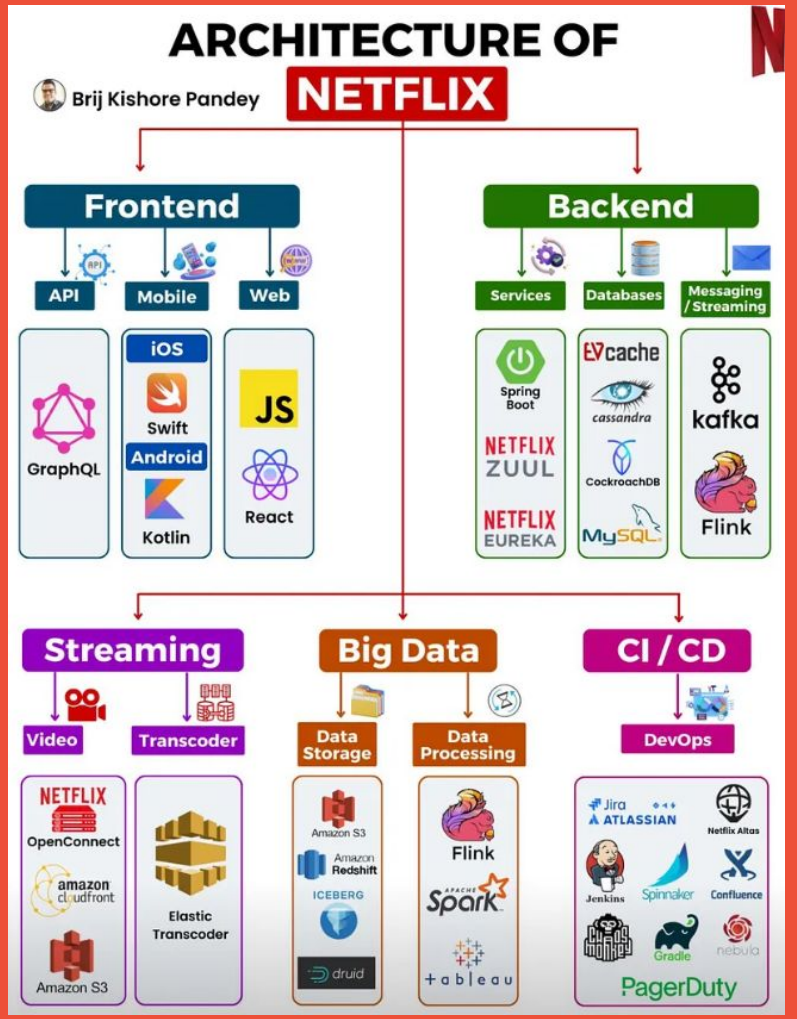


Big Data: O **Amazon S3** e o **Amazon Redshift** formam a base do armazenamento e **warehousing de dados** da Netflix. O **ICEBERG** e o **Druid**, poderosos sistemas de dados distribuídos, aprimoram ainda mais os recursos de **processamento de dados**. O **Spark** e o **Flink**, como mecanismos de **processamento de dados**, processam **grandes quantidades de dados**, enquanto o Tableau capacita a **visualização de dados para insights valiosos.**

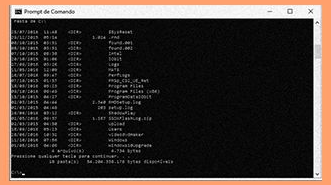


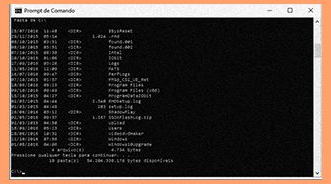
Streaming: Na Netflix, **Kafka e Flink** são a dupla dinâmica por trás do **processamento de dados em tempo real e da entrega de conteúdo**.

O **Kafka gerencia fluxos de dados de forma eficiente**, enquanto o poderoso mecanismo do **Flink processa dados em tempo real**, garantindo uma experiência de streaming perfeita para milhões de usuários no mundo todo. Juntos, eles permitem que a Netflix fique à frente no mundo em constante evolução dos serviços de streaming.



programa{m}aria
/sprint





Arquitetura de Dados: Desafios atuais

Segurança e Privacidade de Dados

Proteção contra ameaças: Com o aumento das violações de dados, proteger informações sensíveis é crucial.

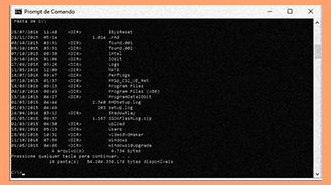
Conformidade regulatória: Atender a regulamentações como GDPR e LGPD exige esforços contínuos para garantir a privacidade dos dados.

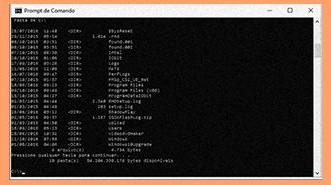
Escalabilidade e Perfomance

Crescimento de dados: Gerenciar grandes volumes de dados sem comprometer a performance é um desafio constante.

Infraestrutura: Garantir que a infraestrutura suporte o crescimento e a demanda de processamento é essencial

programa{m}aria
/sprint





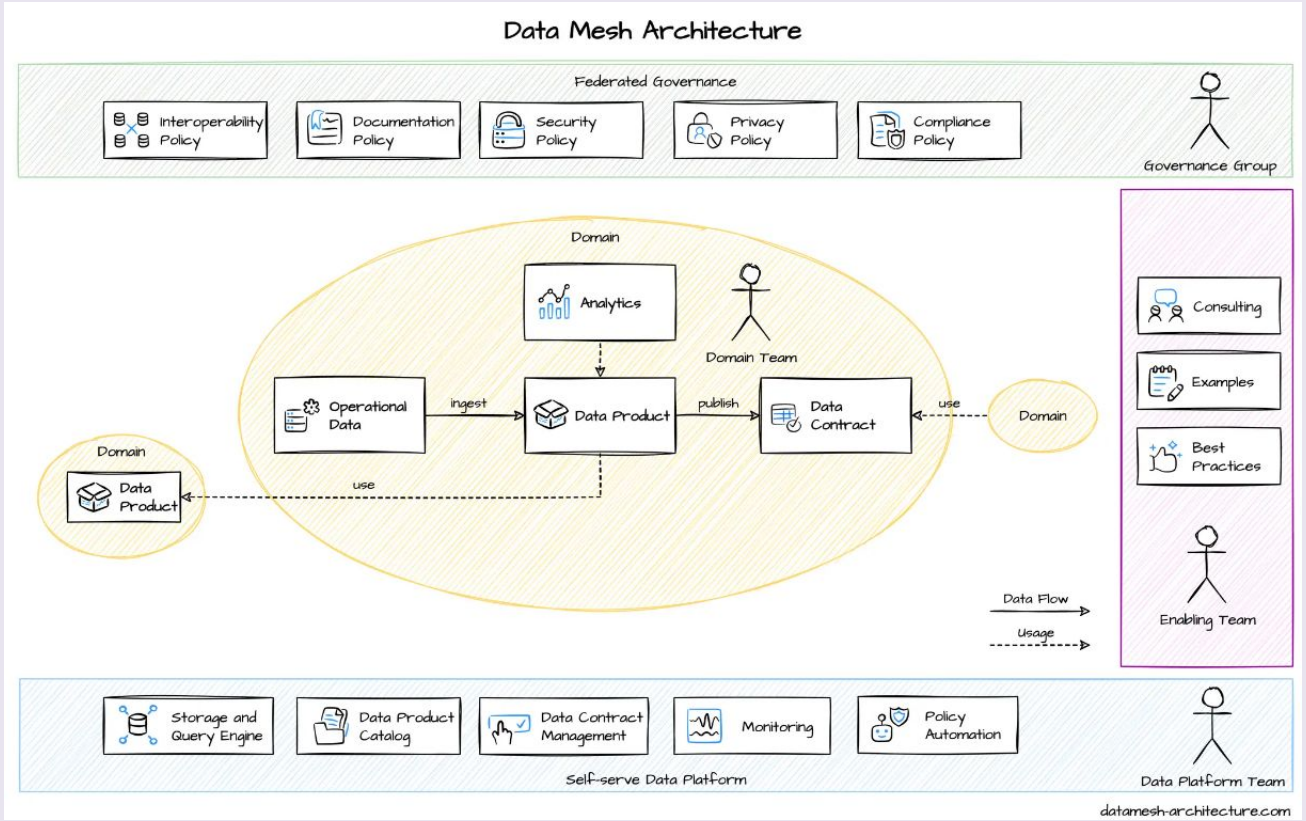
Arquitetura de Dados: Tendências

Data Fabric e Data Mesh

Data Mesh

Data Mesh é uma abordagem moderna e descentralizada para a gestão de dados analíticos. Em vez de centralizar todos os dados em um único data lake ou data warehouse, o Data Mesh distribui a propriedade dos dados para equipes específicas de domínio, que gerenciam e fornecem os dados como um produto. Isso permite que os usuários finais acessem e consultem os dados diretamente onde eles estão, sem a necessidade de transportá-los para um local central

Data Mesh


Data Mesh Architecture
Federated Governance
Interoperability Policy
Documentation Policy
Security Policy
Privacy Policy
Compliance Policy
Governance Group
Domain
Analytics
Domain Team
Operational Data
ingest
Data Product
publish
Data Contract
use
Domain
Domain
Data Product
use
Consulting
Examples
Best Practices
Enabling Team
Data Flow
Usage
Storage and Query Engine
Data Product Catalog
Data Contract Management
Monitoring
Policy Automation
Self-serve Data Platform
Data Platform Team
datamesh-architecture.com

Data Mesh

Benefícios do Data Mesh:

#1 **Descentralização da Propriedade dos Dados**

- Autonomia das equipes de domínio na gestão e análise dos dados.
- Maior responsabilidade e qualidade dos dados.

#2 **Escalabilidade Organizacional**

- Crescimento sustentável da infraestrutura de dados.
- Flexibilidade para adaptar práticas de gestão de dados às necessidades específicas.

#3 **Melhoria na Qualidade dos Dados**

- Tratamento dos dados como um produto, focando na qualidade e usabilidade.
- Ciclo de feedback contínuo entre produtores e consumidores de dados.

#4 **Governança Federada**

- Consistência e padronização dos dados entre diferentes domínios.
- Aplicação de políticas de conformidade e segurança de forma federada.

Data Fabric

Data Fabric é uma arquitetura de gerenciamento de dados que permite a integração e o gerenciamento contínuos de dados em diversos ambientes. Em vez de centralizar os dados em um único local, o Data Fabric conecta virtualmente diferentes fontes de dados, facilitando o acesso e a análise sem a necessidade de cópias redundantes[1][2].

Data Fabric

Benefícios do Data Fabric:

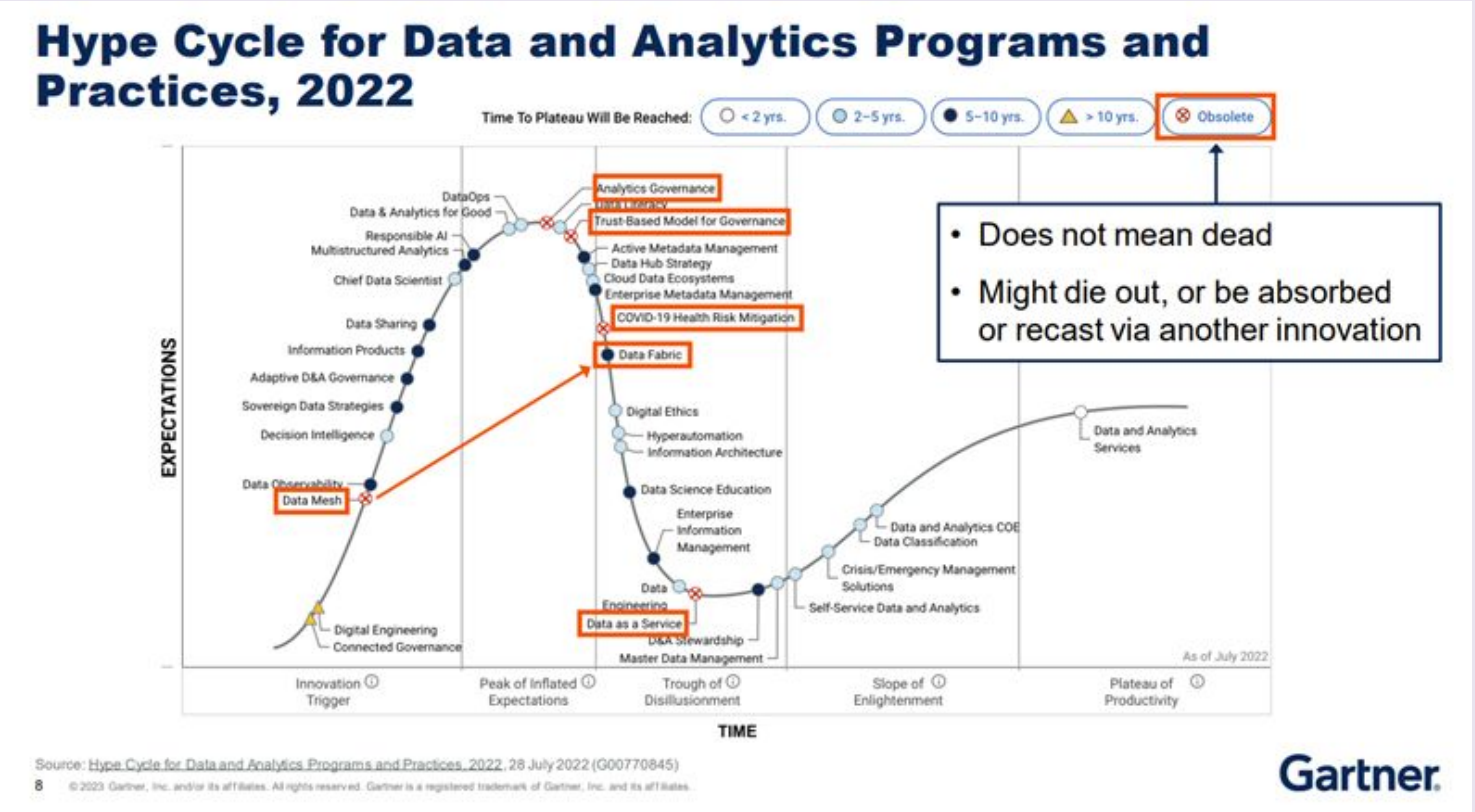
#1 **Integração Simplificada**: Facilita a integração de dados de várias fontes, sejam elas locais, na nuvem ou híbridas.

#2 **Acesso em Tempo Real**: Permite que os usuários acessem dados em tempo real, melhorando a tomada de decisões e a eficiência operacional.

#3 **Governança Automatizada**: Aplica políticas e regras de governança de dados de forma automatizada, garantindo conformidade e segurança.

#4 **Flexibilidade**: Oferece flexibilidade para integrar APIs e outras plataformas, adaptando-se às necessidades específicas de cada organização.

Data Fabric e Data Mesh



Source: Hype Cycle for Data and Analytics Programs and Practices, 2022, 28 July 2022 (G00770845)

Gartner

Is data mesh dead?

Desafios enfrentados na adoção do data mesh, segundo o levantamento do Gartner:

- Muitas empresas têm sua própria definição de data mesh.
- Exige muito esforço de consultoria.
- A equipe de implementação precisa ser qualificada em produtos de dados (ele/ela precisa saber como operacionalizar o uso de dados, quando desativá-los, etc.).
- 80% dos entrevistados em uma pesquisa disseram que não estavam prontos para serem responsabilizados pela governança.

Em um mundo dominado pelos Dados e pela Inteligência Artificial, a arquitetura de dados não é apenas uma necessidade técnica, mas um diferencial estratégico.

Investir em uma arquitetura de dados sólida é crucial para qualquer organização que deseja se manter competitiva e inovadora!

Ainda não se inscreveu para o evento? Inscreva-se para receber todos os conteúdos da Sprint e garantir seu certificado!

Link em breve

Muito obrigada!

linkedin: https://www.linkedin.com/in/deboramariano/