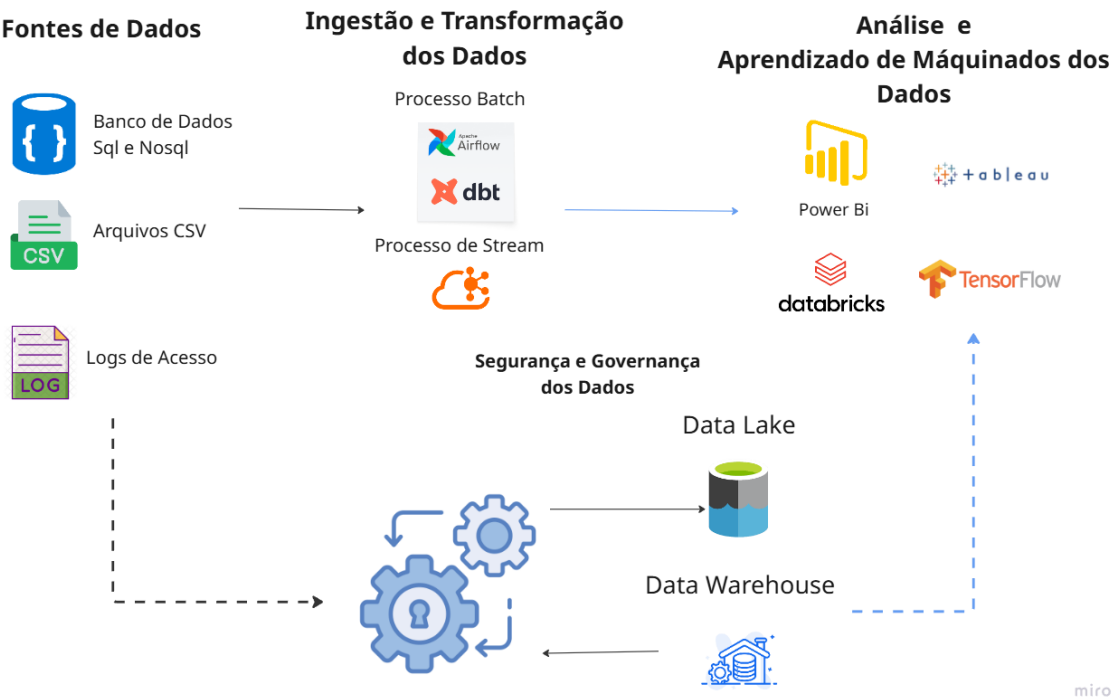
Projeto de Arquitetura de Dados para E-commerce

1. Introdução

Este projeto tem como objetivo desenvolver uma arquitetura moderna de dados para uma empresa de e-commerce em expansão, centralizando informações dispersas em múltiplas fontes (SQL, NoSQL, arquivos CSV, e logs de acesso). A solução proposta deve garantir escalabilidade, governança de dados, segurança e suporte a análises avançadas, com possibilidades futuras para a inclusão de machine learning.

2. Descrição da Arquitetura

Diagrama Geral:



Componentes:

1. **Fontes de Dados**
    - Banco SQL: MySQL/PostgreSQL (dados transacionais).
    - Banco NoSQL: MongoDB (dados de navegação, comportamento).
    - CSVs: Exportações de sistemas legados.
    - Logs: Servidores locais com arquivos de log.
2. **Ingestão de Dados (ETL/ELT)**
    - Lote: Apache Airflow + dbt (transformações SQL para Data Warehouse).
    - Streaming: Apache Kafka para ingestão de logs em tempo real.

3. **Armazenamento**
    - **Data Lake (S3/GCS/ADLS):** Nível Bronze (dados brutos), Silver (dados tratados), Gold (dados prontos para consumo analítico).
    - **Data Warehouse (BigQuery/Redshift/Snowflake):** Dados modelados em esquema estrela para BI e análises gerenciais.
4. **Camada Analítica**
    - BI: Power BI ou Looker para dashboards.
    - Machine Learning: Notebooks em Databricks ou Vertex AI (GCP).
5. **Governança e Segurança**
    - IAM com RBAC (papéis por função).
    - Criptografia em repouso e em trânsito.
    - Data Catalog e lineage com Google Data Catalog ou AWS Glue Data Catalog.

3. Justificativa das Escolhas

- **Data Lake + Data Warehouse:** Permite flexibilidade para dados estruturados e não estruturados, e otimiza custo/desempenho.

  Data Lake (S3 ou Google Storage):

    - Baixo custo para armazenamento de grandes volumes de dados não estruturados
    - Suporte nativo a logs e arquivos CSV
    - Camadas bronze (raw), silver (curated), gold (trusted)

  BigQuery (ou Snowflake):

    - Armazena dados analíticos de forma escalável
    - Bom desempenho em SQL.
    - Modelo de precificação “pas as you go”, ou seja, você paga pelo que usar.

- **Airflow + dbt:** Airflow para orquestração e dbt para transformação modular e reutilizável.
- **Kafka:** Ideal para ingestão contínua de logs com alta taxa de eventos.
- **Power BI/Tableau:** Interface amigável e integração com modelos de dados do warehouse.
- **Databricks ou Vertex AI:** Suporte a notebooks e modelos preditivos com integração direta ao Data Lake.

- **Tensor Flow:** O framework de machine learning de código aberto oferece boa compatibilidade arquitetônica e facilita a implantação de frameworks computacionais em diversas plataformas.

4. Modelo de Dados

A modelagem adotada para o Data Warehouse segue o **esquema estrela**, uma abordagem eficiente para análises e relatórios, com uma tabela fato central ligada a múltiplas tabelas dimensão.

Data Warehouse (Camada Analítica):

Tabela Fato: fato_pedidos

- Contém os registros transacionais de pedidos realizados na plataforma.
- Principais atributos:
    - id_pedido
    - id_cliente
    - id_produto
    - id_tempo
    - id_categoria
    - quantidade
    - valor_total
    - desconto_aplicado
    - forma_pagamento
    - canal_venda (web, app, marketplace)

Tabelas Dimensão:

- **dim_clientes**: dados demográficos, localização, comportamento de compra.
    - id_cliente, nome, email, idade, sexo, cidade, estado, score_fidelidade
- **dim_produtos**: informações dos produtos vendidos.
    - id_produto, nome_produto, marca, preço_base, status_estoque
- **dim_tempo**: detalhamento temporal para análise ao longo do tempo.
    - id_tempo, data, dia, mes, ano, trimestre, semana, dia_da_semana

- **dim_categorias**: classificação dos produtos em categorias e subcategorias.
    - id_categoria, categoria, subcategoria, segmento

Data Lake:

- **Bronze:** CSVs e logs brutos, dumps NoSQL.
- **Silver:** Dados tratados e padronizados.
- **Gold:** Dados prontos para análise, compatíveis com o Data Warehouse.

Integração:

- MongoDB e logs estruturados via Kafka são normalizados no Silver e integrados ao Gold.

5. Plano de Governança e Segurança

- **Controle de Acesso:**
    - IAM com RBAC e ABAC.
    - Integração com Active Directory/SSO.
- **Proteção de Dados Sensíveis:**
    - Criptografia (AES-256 em repouso, TLS em trânsito).
    - Mascaramento e anonimização (colunas de CPF, e-mail).
    - Compliance com LGPD e GDPR.
- **Monitoramento e Auditoria:**
    - Logging centralizado via Stackdriver ou CloudWatch.
    - Alertas de acesso não autorizado.
    - Auditoria de pipelines com trilhas de execução (Airflow logs).
- **Qualidade dos Dados:**
    - Data Catalog para metadata.
    - Validadores automáticos (ex: Great Expectations).
    - Versionamento e lineage.

---

6. Conclusão

A arquitetura proposta atende às necessidades atuais e futuras da empresa, proporcionando uma base robusta para análises em tempo real, inteligência de negócios e aprendizado de máquina. A integração entre Data Lake e Data Warehouse garante flexibilidade e desempenho, enquanto práticas de segurança e governança asseguram conformidade e proteção dos dados sensíveis, seguindo as normas e orientações da LGPD.