Engenharia de Dados

1. Pipeline de Dados

O **pipeline de dados** é o processo que automatiza a movimentação de dados entre diferentes sistemas, garantindo sua coleta, transformação e armazenamento de forma contínua.

Etapas de um Pipeline:

- **Fonte de Dados:** Sistemas de onde os dados são coletados.
- **ETL/ELT:** Processo de transformação dos dados.
- **Destino Final:** Data Warehouse ou Data Lake, onde os dados ficam disponíveis para análise.

---

2. ETL vs ELT

ETL (Extract, Transform, Load): Os dados são extraídos, transformados e depois carregados no destino.
ELT (Extract, Load, Transform): Os dados são extraídos e carregados diretamente no destino, e a transformação ocorre posteriormente.

Diferença Principal:

- **ETL:** Adequado para data warehouses tradicionais.
- **ELT:** Melhor para data lakes e ambientes modernos.

---

3. Data Warehousing

Um **data warehouse** é um repositório centralizado que organiza e armazena dados estruturados para consultas rápidas e análise de dados.

Principais características:

- Dados estruturados.
- Focado em análise histórica e relatórios.
- Exemplo de ferramentas: Snowflake, Amazon Redshift, Google BigQuery.

4. Data Lake

Um **data lake** armazena dados em seu formato original (estruturado, semiestruturado e não estruturado). É mais flexível, mas requer governança adequada para evitar um "data swamp" (pântano de dados).

Diferença entre Data Lake e Data Warehouse:

- **Data Warehouse:** Dados estruturados, organizados para consultas.
- **Data Lake:** Dados brutos, sem estrutura definida.

---

5. Big Data

Big Data refere-se a conjuntos massivos de dados que requerem tecnologias especiais para armazenamento e processamento.

Os 5 Vs do Big Data:

- **Volume:** Grande quantidade de dados.
- **Velocidade:** Dados gerados em alta frequência.
- **Variedade:** Diferentes formatos de dados.
- **Veracidade:** Dados confiáveis e precisos.
- **Valor:** Insights obtidos dos dados.

---

6. Armazenamento de Dados

Armazenamento de dados envolve escolher a melhor tecnologia para guardar informações de forma segura e acessível.

Principais tipos de armazenamento:

- **Bancos de Dados Relacionais (SQL).**
- **Bancos de Dados NoSQL.**
- **Armazenamento em Nuvem:** AWS S3, Azure Blob Storage, Google Cloud Storage.

---

7. Tecnologias e Ferramentas

Ferramentas comuns na engenharia de dados:

- **ETL:** Apache NiFi, Talend, Informatica.
- **Big Data:** Hadoop, Spark, Kafka.
- **Data Warehouse:** Snowflake, Redshift, BigQuery.
- **Data Lake:** AWS S3, Azure Data Lake, Google Cloud Storage.

---

8. Qualidade de Dados

A **qualidade de dados** refere-se à precisão, consistência e integridade das informações.

Principais aspectos:

- **Validade:** Dados estão no formato correto?
- **Consistência:** Dados coerentes entre sistemas.
- **Integridade:** Dados completos e sem lacunas.

---

9. Arquitetura de Dados

A **arquitetura de dados** define como os dados são coletados, armazenados e processados dentro de uma organização.

Principais componentes:

- **Fonte de Dados:** Origem das informações.
- **Pipeline de Dados:** Fluxo de processamento.
- **Armazenamento:** Data lake ou warehouse.
- **Ferramentas de análise:** Power BI, Tableau, etc.

---

10. Arquitetura Serverless

Na **arquitetura serverless**, o desenvolvedor não precisa gerenciar servidores diretamente. A infraestrutura é escalada automaticamente.

Benefícios:

- Menos custos operacionais.
- Escalabilidade automática.
- Foco no código e nas funcionalidades.