**Projeto: Pipeline de ETL (Extração, Transformação e Carga) para Análise de Dados de Vendas**

**Objetivo**: Desenvolver um pipeline de ETL automatizado para coletar, processar, transformar e armazenar dados de vendas, preparando-os para análise e relatórios.

**Etapas do Projeto:**

1. **Coleta de Dados (Extração):**
   * **Fontes de Dados**:
     + Coleta de dados de diferentes fontes, como APIs, arquivos CSV, bancos de dados SQL, e fontes externas como Google Sheets ou arquivos Excel.
   * **Integração de Dados**:
     + Configure scripts para coletar dados automaticamente, com agendamentos usando cron jobs ou ferramentas como Airflow.
2. **Processamento e Limpeza de Dados (Transformação):**
   * **Limpeza de Dados**:
     + Trate valores ausentes e dados duplicados.
     + Corrija formatos de dados inconsistentes e erros.
   * **Transformação de Dados**:
     + Realize agregações, cálculos de novas colunas (como margens de lucro), e aplique regras de negócios.
     + Normalize os dados para uma estrutura consistente.
   * **Enriquecimento de Dados**:
     + Adicione novos dados contextuais, como informações demográficas de clientes, ou taxas de câmbio para conversões de moeda.
   * **Validação de Dados**:
     + Implemente validações para garantir que os dados processados atendam às expectativas e regras definidas.
3. **Carga de Dados (Carga):**
   * **Armazenamento de Dados**:
     + Salve os dados transformados em um banco de dados, como MySQL, PostgreSQL, ou em um data warehouse como Amazon Redshift ou Google BigQuery.
   * **Otimização do Armazenamento**:
     + Use técnicas de particionamento e indexação para melhorar a performance das consultas.
   * **Automação da Carga**:
     + Configure jobs automatizados para a carga de dados, utilizando ferramentas como **Apache Airflow** ou scripts agendados.
4. **Análise e Visualização de Dados:**
   * **Exploração de Dados**:
     + Utilize **Pandas** para análise exploratória de dados (EDA), criando tabelas dinâmicas e resumos.
   * **Visualização**:
     + Use **Matplotlib**, **Seaborn**, ou **Tableau** para criar dashboards que mostrem insights sobre vendas, como tendências de vendas, produtos mais vendidos, e desempenho por região.
   * **Relatórios Automatizados**:
     + Gere relatórios periódicos em PDF ou Excel, automatizando a criação e envio via e-mail usando smtplib ou ferramentas como **Pandas-Reports**.
5. **Documentação e Monitoramento:**
   * **Documentação do Pipeline**:
     + Documente cada etapa do pipeline, explicando o fluxo de dados, as transformações aplicadas e as regras de negócios implementadas.
   * **Monitoramento**:
     + Implemente logging e alertas para monitorar falhas no pipeline e garantir a integridade dos dados.
   * **Testes de Qualidade de Dados**:
     + Desenvolva scripts para testar a qualidade dos dados em cada etapa do processo ETL, garantindo que os dados carregados estejam corretos.
6. **Deploy e Manutenção (Opcional):**
   * Implemente o pipeline em um ambiente de produção, utilizando servidores locais ou plataformas em nuvem como AWS, GCP ou Azure.
   * Configure tarefas de manutenção, como a limpeza de dados antigos ou o arquivamento de dados históricos.

**Tecnologias e Bibliotecas Usadas:**

* **Python**: Linguagem principal.
* **Pandas**: Manipulação e transformação de dados.
* **SQLAlchemy**: Integração com bancos de dados SQL.
* **Apache Airflow**: Orquestração e automação do pipeline.
* **MySQL/PostgreSQL/Redshift/BigQuery**: Armazenamento de dados.
* **Matplotlib/Seaborn/Tableau**: Visualização de dados.
* **smtplib**: Envio automatizado de relatórios por e-mail.
* **AWS/GCP/Azure**: Implementação e escalabilidade na nuvem.

**Resultado Esperado:**

Ao final deste projeto, você terá construído um pipeline de ETL robusto que é capaz de lidar com grandes volumes de dados de vendas, garantindo que eles sejam limpos, transformados e armazenados de maneira eficiente para análise. Esse projeto demonstra habilidades em manipulação de dados, automação, integração com várias fontes de dados e criação de pipelines escaláveis, o que é altamente valorizado em qualquer função relacionada a ciência de dados, análise de dados ou engenharia de dados.