Desenvolvimento de um Pipeline ETL/ELT para Geração de Dados em Parquet

Objetivo:

O objetivo deste projeto é que os alunos, trabalhando em grupos, demonstrem a capacidade de construir um pipeline de extração, transformação e carga (ETL) ou extração e carga e transformação (ELT) completo, desde a obtenção dos dados até a geração de um arquivo final em formato Parquet.

O projeto visa consolidar os conhecimentos adquiridos durante o curso sobre as técnicas de extração de dados, tratamento e preparação de dados, e armazenamento de dados em formatos otimizados para análise.

Descrição:

Os grupos deverão escolher (uma ou mais) fonte(s) de dados de sua preferência (banco de dados relacional, arquivos CSV, APIs, etc.) e definir um conjunto de dados a ser extraído. Em seguida, os alunos deverão:

1. Definição do processo de ETL/ELT
2. Extração dos Dados
3. Transformação:
    a) Integração
    b) Limpeza
    c) Padronização
    d) Enriquecimento
    e) Filtros e Seleções
    f) f)...
4. Gerar arquivo final em .parquet

Observações:

- Os grupos são livres para escolher a(s) fonte(s) de dados, as ferramentas e as tecnologias a serem utilizadas
- É importante que os alunos demonstrem capacidade de trabalhar em equipe e resolver problemas
- Não será cobrada visualização de dados, mas caso o grupo queira utilizar alguma ferramenta para mostrar resultado de análises na base final, podem desenvolver em qualquer ferramenta.