Bootcamp: Engenheiro(a) de Dados

Trabalho Prático

Módulo 1: Fundamentos em Engenharia de Dados

Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

1. Fundamentos de Engenharia de Dados.
2. Pipeline de Dados.
3. Arquiteturas de Big Data.
4. Metodologia Ágil na Engenharia de Dados.

Enunciado

Você foi contratado por uma consultoria de dados para atender um cliente que construirá seu ambiente de Big Data.

Sua equipe possui pouca experiência e caberá a você, Engenheiro(a) de Dados, prover as devidas orientações técnicas, bem como ajudar na construção de sistemas, considerando as melhores práticas.

Serão necessárias algumas práticas, incluindo extração de dados públicos do Portal da Transparência do Governo Federal:

- https://portaldatransparencia.gov.br/

Vamos utilizar os Dados de Benefícios ao Cidadão, Novo Bolsa Família, os meses de janeiro, fevereiro e março de 2024 (https://portaldatransparencia.gov.br/download-de-dados/novo-bolsa-familia).

Para melhor compreensão dos dados, o dicionário de dados poderá ser utilizado:

- Dicionário de Dados - Novo Bolsa Família (portaldatransparencia.gov.br)

Para a extração, poderão ser utilizadas ferramentas para construção de pipelines de dados, como o Python (Jupyter) ou Databricks Workflows ou alguma ferramenta de seu conhecimento (conforme exemplos vistos em aula). (ATENÇÃO: os arquivos são grandes [2GB] e não abrem no Excel).

Algumas sugestões de construção com o Python:

- Importar as bibliotecas do Pandas, do ZipFile, requests e io do BytesIO.
- Baixar o arquivo da URL correta.
- Extraia o arquivo com ZipFile e BytesIO.
- E leia o arquivo com o Pandas (read_csv).
- Faça os filtros necessários para responder as questões 2, 3 e 4.
- Salve os dados (opcional no banco de dados de sua preferência) (use as bibliotecas do sqlalchemy e do psycopg2 [para o PostgreSQL]).

Algumas sugestões de construção com o Databricks:

- Abra sua conta do Databricks.
- Crie seu workspace.
- Crie o cluster das máquinas.
- Adicione seu Notebook para a importação dos dados.
- Importe as bibliotecas necessárias (requests, Spark, ZipFile).

- Renomeie as colunas (se necessário).
- Crie a tabela.
- Use as consultas com %sql para responder às questões 2, 3 e 4.

Realizada a extração, deve-se responder aos seguintes questionamentos acerca dos dados:

1) Como os dados estão dispostos nos arquivos?

2) Qual o melhor tipo de banco de dados para se armazenar?

3) Como será o fluxo de ingestão desses dados?

4) Qual a frequência de atualização?

5) Qual o método de atualizações será utilizado?

Após a extração, o aluno deverá completar a pipeline, armazenando os dados em um banco de dados relacional. A escolha aqui será individual, mas recomenda-se o uso do PostgreSQL ou do MySQL.

DIVIRTA-SE!