Bootcamp IGTI: Engenheiro de Dados

Desafio

Módulo 4	**Pipeline de dados**

Objetivos

- ✓ Desenvolvimento de crawlers;
- ✓ Pipeline de ETL;
- ✓ Orquestração e automatização de DataFlow;
- ✓ Análise Exploratória de Dados (EDA).

Enunciado

Você é uma pessoa da área de Engenheira de Dados em uma empresa de marketing. Um de seus principais clientes é o Disney Plus, um novo serviço de Streaming disponível no Brasil. Com a forte concorrência pelo mercado de streaming (Netflix, HBOGO, Globoplay etc.), a empresa deseja saber o que os clientes do Brasil estão comentando no Twitter sobre o seu Streaming e sobre seus concorrentes.

Para atender a essa exigente demanda, você precisa desenvolver um *crawler* que fará a captura em tempo real de tweets contendo palavras-chave relacionadas aos serviços de streaming citados. Use a criatividade para escolher as palavras-chave!

Depois de coletar dados por algum tempo, você deve implementar um pipeline automatizado para fazer a limpeza, estruturação e organização dos dados e, por fim, depositá-los em seu DW (para os fins desta atividade, uma tabela relacional no SGBD de sua preferência). Lembre-se de que o pipeline deve ser completamente autônomo e não deve ter intervenção humana para sua execução.

Boa sorte e divirta-se! *Happy coding*!

Atividades

Você deverá desempenhar as seguintes atividades:

1. Escrever um *crawler* para captura de dados em tempo real no Twitter;
2. Após algum tempo de execução, implementar um pipeline automatizado utilizando a ferramenta de sua preferência (Airflow, Prefect, Nifi, Pentaho etc.) para limpeza, organização e estruturação dos dados;
3. Escrever os dados limpos e tratados em uma tabela relacional em SGBD de sua escolha.