Infraestrutura do Projeto

Primeira parte do pipeline é a obtenção dos dados via web scraping: extrair, organizar e guarda em um xlsx. Ou seja, o ETL. Note que não é um processo automatizado real-time, pois o script precisa ser acionado manualmente e os dados não estão sendo armazenados em um Sistema Gerenciador de Banco de Dados (SGBD) via SQL (padrão tabular). Mas sim locamente em um arquivo excel.

Idealmente, o processo do pipeline deve ser extrair via web scraping, armazenar em um SGBD, automatizar o script de coleta com algum orquestrador – possivelmente o Apache Airflow (deixando o script rodar 1 vez ao mês, por exemplo. Assim, atualiza os dados evitando defasagem, perda de informações sazonais ou viés amostral).

A terceira etapa do processo do pipeline é o pré-processamento de dados: organização, limpeza, filtragem e compatibilização entre as diversas fontes de coleta.

→ Mais etapas a definir até chegar na modelagem