Autoavaliação

O objetivo desse projeto era construir um pipeline completo de Engenharia de Dados, utilizando o Databricks Free Version, para analisar o desmatamento no Brasil a partir de dados públicos do **MapBiomas**. Para isso, pretendia desenvolver todas as etapas de ingestão, limpeza, transformação e modelagem dos dados até a criação de um dashboard analítico capaz de responder às perguntas formuladas no início do projeto.

Considero que os principais objetivos foram alcançados. O pipeline foi implementado seguindo a arquitetura medalhão (Bronze, Silver e Gold), proposta durante as aulas, linguagens SQL, Python e PySpark. O resultado foi um modelo estrela funcional e bem estruturado. Além disso, realizei a integração de dados do IBGE (2024) para o cálculo de proporções de área desmatada, que trouxe uma camada analítica mais robusta e interpretável. O dashboard final é totalmente interativo e permite muitos insights sobre o tema.

Durante o desenvolvimento do projeto, no entanto, enfrentei algumas dificuldades. Uma delas foi a tentativa de incorporar as áreas dos municípios para calcular os percentuais de desmatamento local. Essa abordagem apresentou inconsistências, pois várias proporções de desmatamento ultrapassavam 100%. Concluí que os dados eram incompatíveis devido às mudanças históricas de fronteiras que inviabilizaram a comparação temporal. Por isso, optei por restringir a análise proporcional apenas ao nível estadual, cujas fronteiras são estáveis.

Outra dificuldade foi a necessidade de substituir o dataset no meio do projeto. Durante a construção do dashboard, notei a ausência dos estados do Amazonas e Amapá no mapa interativo (evidenciando a importância da visualização em um projeto de dados). Essa descoberta me levou a identificar que o arquivo que eu estava utilizando estava incompleto. Assim, voltei ao site do MapBiomas e consegui com facilidade uma versão atualizada e completa do dataset.

Desafios fazem parte de qualquer projeto de dados. O desenvolvimento raramento ocorre de forma linear, sendo necessário revisitar etapas diversas vezes por erros identificados apenas em fases posteriores. É essa dinâmica que possibilita o aprendizado técnico e o desenvolvimento de uma visão crítica a respeito da qualidade e coerência dos dados.

Dessa forma, considero que essa experiência consolidou minha compreensão sobre o ciclo completo de dados em ambiente de nuvem, resultando em um pipeline funcional, bem documentado e consistente.