Documentação do Case Técnico

Introdução

Este case técnico foi desenvolvido com o objetivo de demonstrar habilidades em análise de dados, geração de insights a partir de datasets e aplicação de boas práticas no ciclo de vida dos dados. Aqui, apresentei as etapas desenvolvidas até o momento, detalhando as decisões tomadas e as soluções implementadas. O foco foi na manipulação de dados do Brasileirão, utilizando dados reais e sintéticos para expandir o dataset, além de organizar e catalogar as informações com base em boas práticas de Data Lake.

Etapas Realizadas

Parte 0 - Planejamento

A primeira etapa foi o planejamento do projeto, onde criei um artefato no Notion representando todas as etapas do ciclo de vida dos dados. Utilizei a abordagem ágil e o PMBOK para mapear tarefas, estimar custos e organizar recursos. A partir disso, o projeto seguiu com o cronograma definido, permitindo monitoramento e controle das atividades planejadas.

Parte 1 - Escolha da Base de Dados

Para a escolha da base de dados, optei pelo **Brazilian Soccer Matches Dataset**. Após carregar e analisar os dados, percebi que a quantidade de registros era insuficiente para cumprir os requisitos do teste, que exigia ao menos 100.000 registros. Assim, gerei dados sintéticos para complementar o dataset, totalizando mais de 120.000 registros, mantendo a consistência e o realismo dos dados adicionados.

Parte 2.1 - Integração e Análise Descritiva

Nesta fase, carreguei o dataset no Jupyter Notebook, onde realizei análises descritivas iniciais. Verifiquei a presença de valores ausentes e executei um resumo estatístico das variáveis. Essa etapa foi essencial para garantir que os dados estavam prontos para as análises seguintes e para garantir que os dados sintéticos mantinham a coerência com os registros reais.

Parte 3 - Explorar (Catalogação e Organização dos Dados)

Nesta parte, organizei os dados seguindo a arquitetura de **Data Lake** com as zonas: **Bruta (Raw)**, **Padronizada (Standardized)** e **Curada (Curated)**.

1. **Zona Bruta (Raw)**: Aqui, mantive os dados em seu formato original, tanto os dados reais quanto os dados sintéticos gerados. Essa zona serve como um repositório para os dados brutos, sem modificações ou transformações.
2. **Zona Padronizada (Standardized)**: Na zona padronizada, limpei os dados, tratei valores ausentes e removi duplicidades. Nesta etapa, também foram aplicadas pequenas transformações para padronizar os formatos das colunas, como datas e nomes de times, de forma a facilitar o processamento subsequente.
3. **Zona Curada (Curated)**: Na zona curada, os dados foram preparados para análises avançadas. Aqui, adicionei colunas calculadas, como a diferença de gols entre o time da casa e o time visitante, e a coluna `winner` (vencedor), que identifica o resultado da partida (vitória da casa, vitória do visitante ou empate).

Além disso, cataloguei o dataset utilizando boas práticas de dicionário de dados. Segue o **Dicionário de Dados** gerado:

Coluna	Descrição	Tipo
home_team	Nome do time da casa	String
away_team	Nome do time visitante	String
home_goal	Quantidade de gols do time da casa	Inteiro
away_goal	Quantidade de gols do time visitante	Inteiro
winner	Resultado da partida (Casa, Visitante, Empate)	String
season	Temporada do jogo	Inteiro

home_team_state	Estado do time da casa	String
away_team_state	Estado do time visitante	String

Essa estruturação é fundamental para garantir que as próximas etapas, como análises e geração de relatórios, sejam consistentes e de fácil interpretação.

Parte 7 - Análise de Dados

Realizei a análise de dados utilizando ferramentas como **Matplotlib** e **Seaborn**. Foram criadas cinco visualizações, incluindo:

- **Distribuição de Gols por Partida (Casa vs Visitante)**: Um gráfico que mostra a frequência de gols feitos pelos times da casa e pelos visitantes em cada partida.
- **Proporção de Resultados**: Um gráfico de pizza que exibe a proporção de vitórias da casa, vitórias do visitante e empates.
- **Média de Gols por Time**: Uma análise que mostra a média de gols feitos em casa e fora por cada time.
- **Diferença de Gols por Temporada**: Gráfico de área mostrando como a diferença de gols (Casa - Visitante) evoluiu ao longo das temporadas.
- **Relação entre Gols e Vitórias por Time**: Um gráfico de dispersão que relaciona a quantidade de gols marcados com o número de vitórias de cada time.

Essas visualizações fornecem insights valiosos sobre o desempenho das equipes e o padrão dos resultados ao longo das temporadas.