### ****MVP – Engenharia de Dados****

**Marcos Fernando de Moraes Siliprandi**

**Descrição**

Este trabalho tem como objetivo a construção de um pipeline de dados utilizando tecnologias na nuvem, com foco na plataforma Databricks Community Edition. O pipeline abrangerá todo o ciclo de vida dos dados, incluindo busca, coleta, modelagem, carga e análise. A base de dados utilizada contém informações sobre partidas do Campeonato Brasileiro, estatísticas dos clubes, gols e cartões, permitindo a extração de insights relevantes sobre o desempenho dos times e jogadores ao longo da competição.

### ****Objetivo****

O objetivo deste MVP é estruturar e transformar dados brutos do Campeonato Brasileiro em um formato otimizado para análise, possibilitando a extração de insights que possam responder a perguntas como:

* Quais times apresentam melhor desempenho ao longo das rodadas com base em estatísticas como posse de bola, chutes e passes?
* Há relação entre a quantidade de cartões recebidos e o resultado final da partida?
* Quais jogadores se destacam na artilharia e em quais momentos da partida ocorrem mais gols?
* Como o fator "mandante" influência nos resultados das partidas?
* Existe uma correlação entre estatísticas de jogo e a vitória de um time?

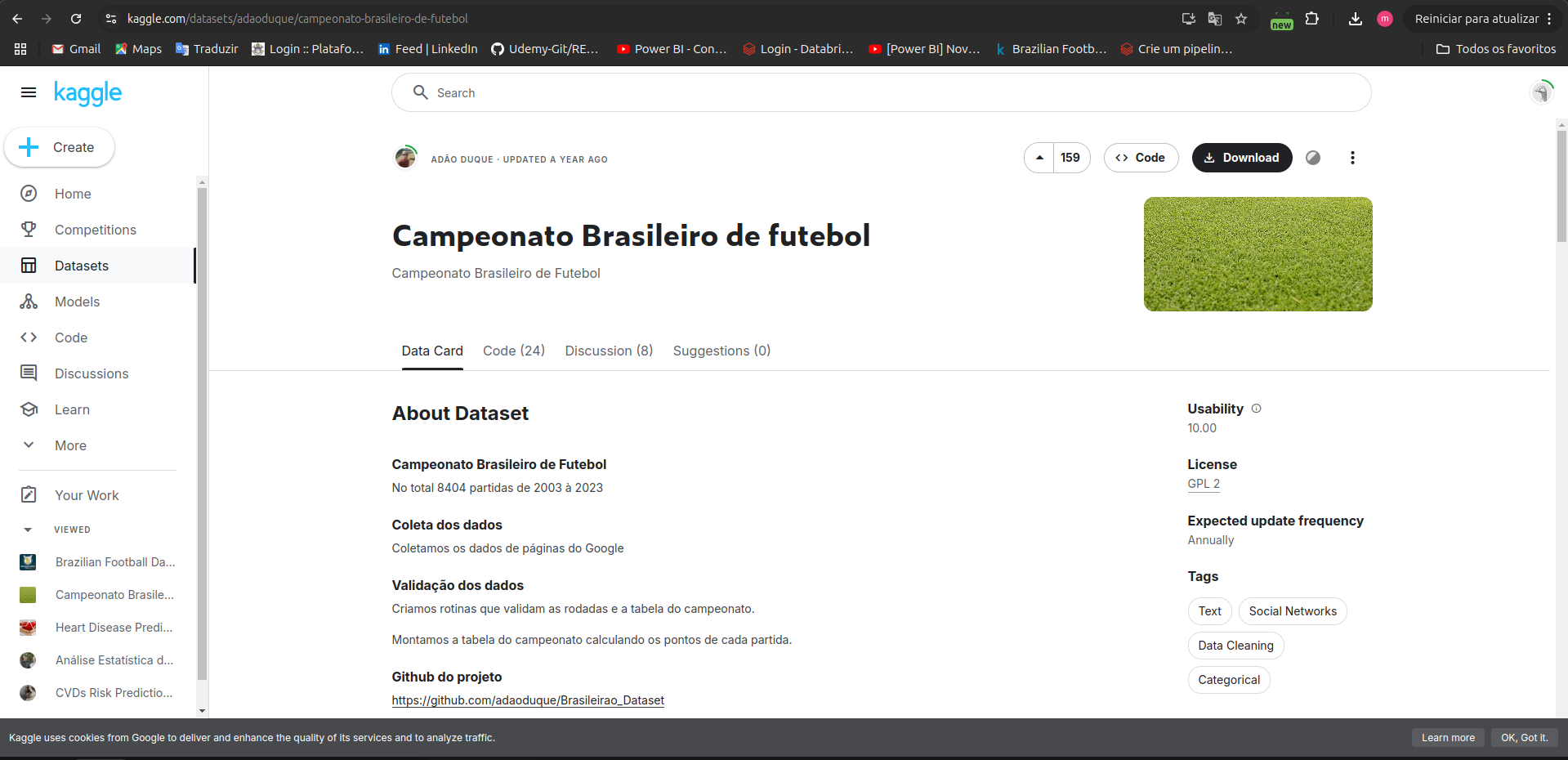
A partir dessa estruturação, será possível garantir a qualidade, consistência e acessibilidade dos dados para futuras análises, facilitando a tomada de decisão baseada em dados.

**Detalhamento**

**1 - Busca pelos dados**

**Para atender aos objetivos deste MVP, optei por utilizar a base de dados "Campeonato Brasileiro de Futebol", disponível gratuitamente no Kaggle. Essa base contém informações detalhadas sobre partidas, estatísticas dos clubes, gols e cartões, permitindo uma análise abrangente do desempenho das equipes ao longo do campeonato.**

A base pode ser acessada através do link:  
[Campeonato Brasileiro de Futebol - Kaggle](https://www.kaggle.com/datasets/adaoduque/campeonato-brasileiro-de-futebol)

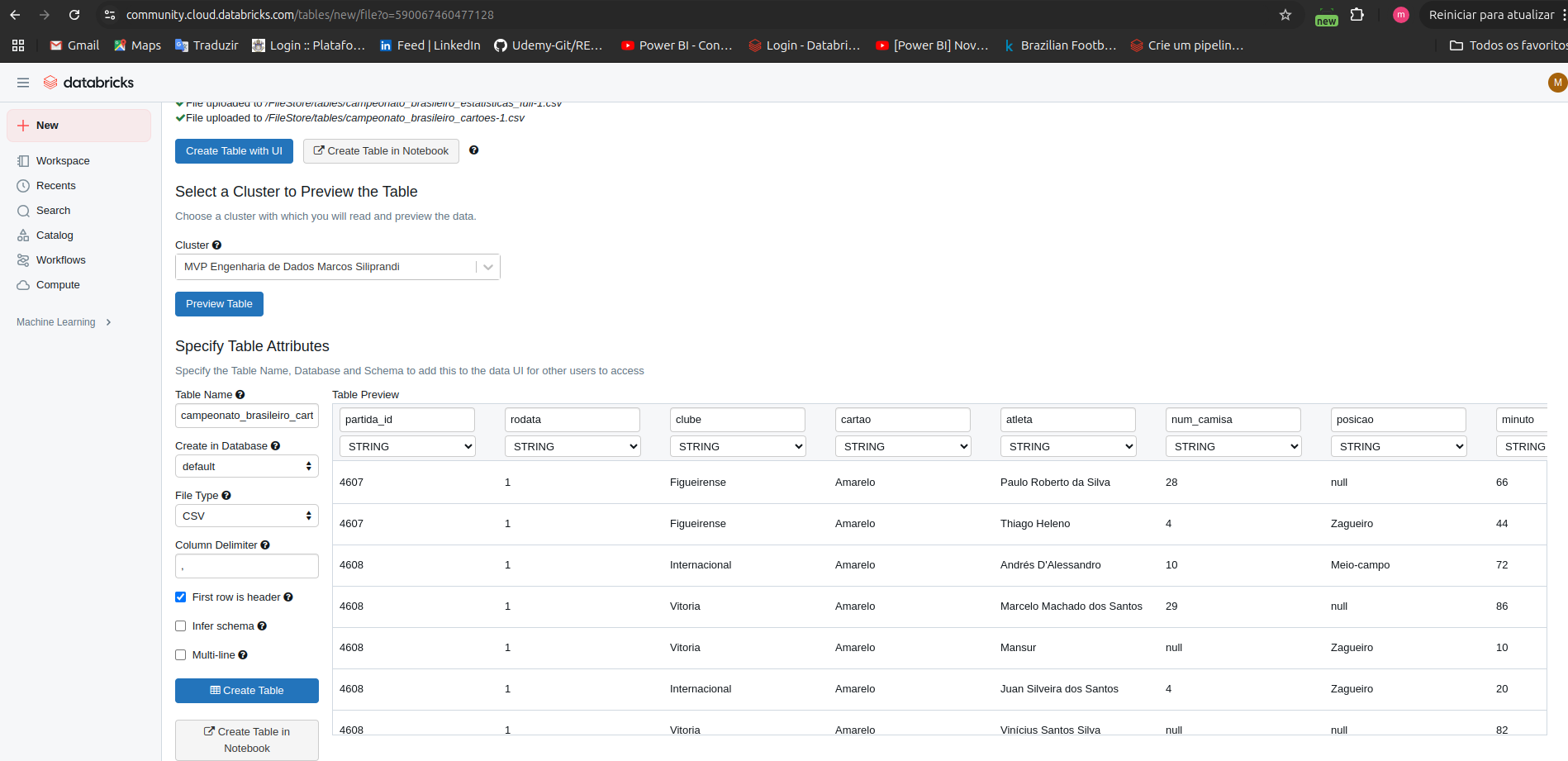
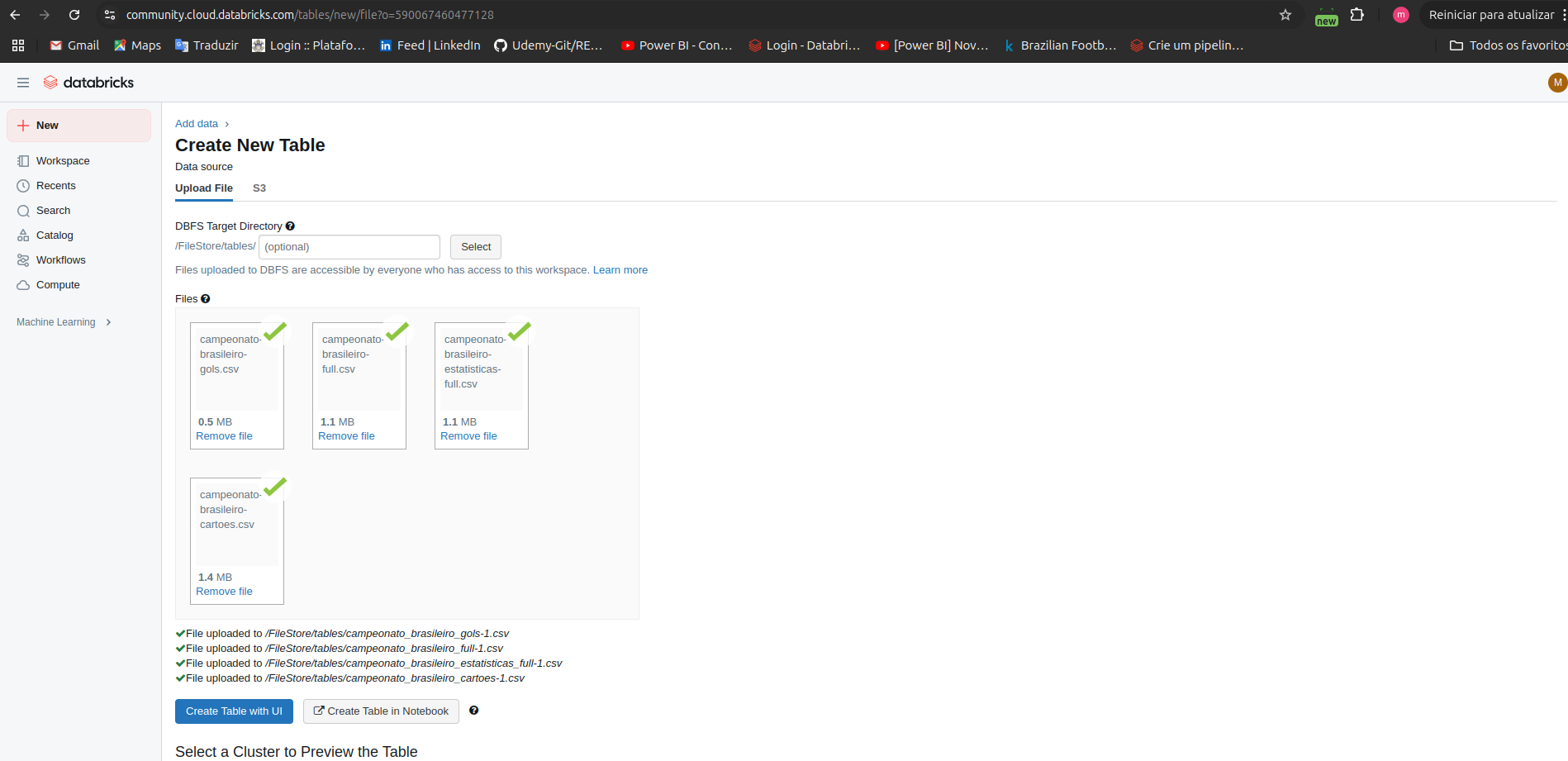


**2 – Coleta dos dados**

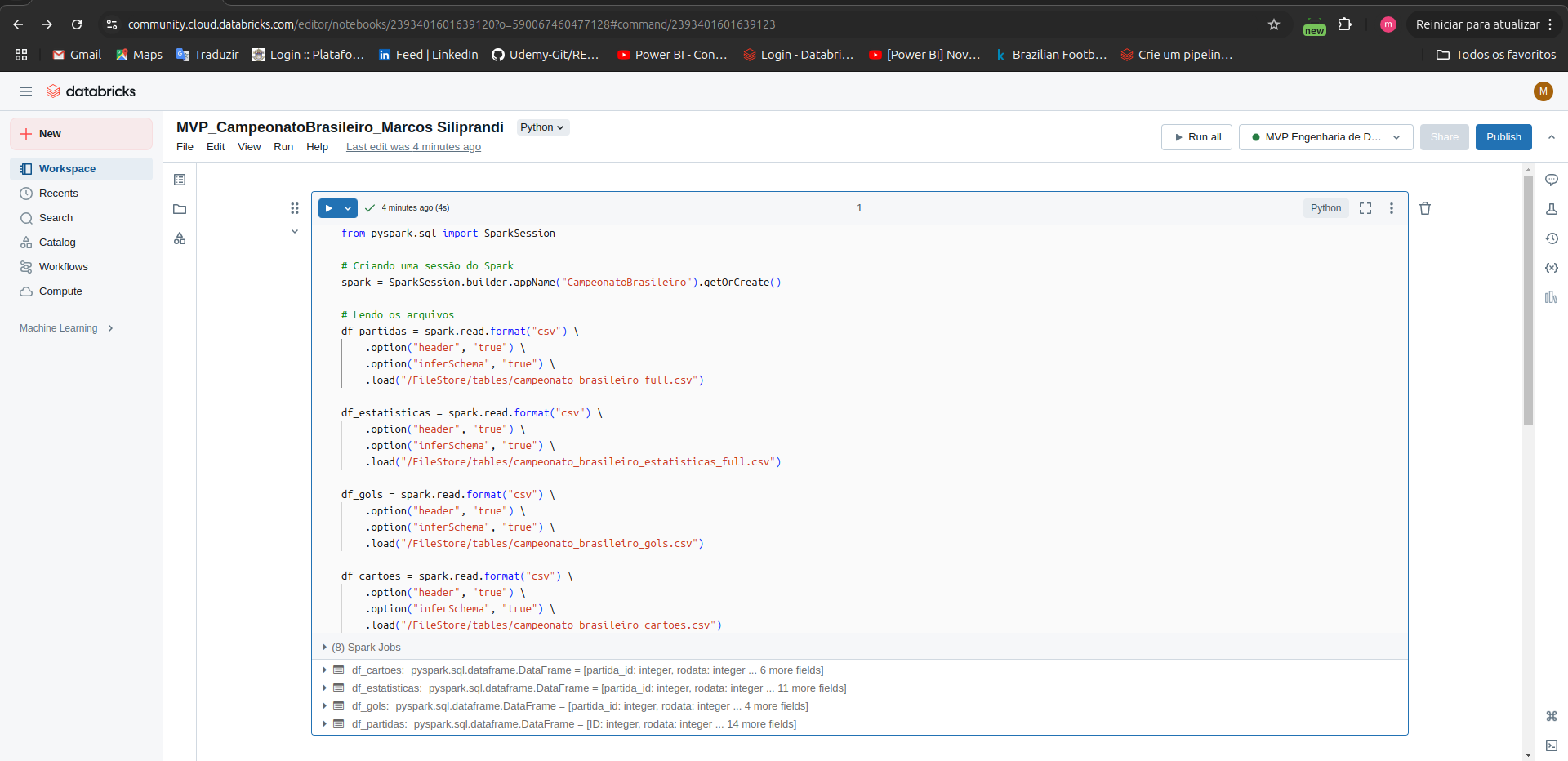
A base de dados escolhida possui 4 arquivos em formato CSV e optei por baixar os arquivos e realizar o upload dos mesmos através do modo gráfico do Databrinks salvando-os no caminho /FileStore/tables/.

Nome dos arquivos:

* campeonato\_brasileiro\_full.csv – Contém os dados das partidas.
* campeonato\_brasileiro\_estatisticas\_full.csv – Contém os dados estatísticos
* campeonato\_brasileiro\_gols.csv – Contém os dados de gols marcados.
* campeonato\_brasileiro\_cartoes.csv – Contém os dados de cartões aplicados.



Em seguida foi criada uma sessão Spark para ler os arquivos CSV e criar os Data frames.



**3 – Modelagem**

Analisando as tabelas, percebi que os dados têm diferentes níveis de detalhe, então faz sentido usar mais de uma tabela fato. Essas tabelas podem se ligar tanto às dimensões quanto entre si, dependendo do caso. Por isso, cheguei à conclusão de que o modelo mais adequado aqui é o **Esquema Estrela com múltiplas tabelas fato**.

**Diagrama

O conteúdo gerado por IA pode estar incorreto.**

**3.1 Catalogo de dados**

**Tabela Fato: F\_Partidas**

|  |  |  |  |
| --- | --- | --- | --- |
| **Coluna** | **Descrição** | **Tipo** | **Domínio** |
| **ID\_Partida** | **Identificador único da partida** | **INT** | **-** |
| **Rodada** | **Rodada do campeonato** | **INT** | **1 - 38** |
| **Data** | **Data do jogo** | **DATE** | **-** |
| **Hora** | **Horário do jogo** | **STRING** | **-** |
| **ID\_Clube\_Mandante** | **Clube mandante (chave estrangeira)** | **INT** | **-** |
| **ID\_Clube\_Visitante** | **Clube visitante (chave estrangeira)** | **INT** | **-** |
| **Formação\_Mandante** | **Formação do time mandante (ex: 4-4-2)** | **STRING** | **-** |
| **Formação\_Visitante** | **Formação do time visitante** | **STRING** | **-** |
| **ID\_Tecnico\_Mandante** | **Técnico do time mandante** | **INT** | **-** |
| **ID\_Tecnico\_Visitante** | **Técnico do time visitante** | **INT** | **-** |
| **ID\_Clube\_Vencedor** | **Clube vencedor da partida** | **INT** | **- / NULL (empate)** |
| **Gols\_Mandante** | **Gols marcados pelo mandante** | **INT** | **0 - 10** |
| **Gols\_Visitante** | **Gols marcados pelo visitante** | **INT** | **0 - 10** |

**Tabela Fato: F\_Cartoes**

|  |  |  |  |
| --- | --- | --- | --- |
| **Coluna** | **Descrição** | **Tipo** | **Domínio** |
| **ID\_cartoes** | **Identificador único do evento de cartão** | **INT** | **-** |
| **Cor** | **Cor do cartão aplicado** | **STRING** | **Amarelo, Vermelho** |
| **ID\_Partida** | **Identificador da partida** | **INT** | **-** |
| **ID\_Clube** | **Clube do atleta penalizado** | **INT** | **-** |
| **ID\_Atleta** | **Atleta penalizado** | **INT** | **-** |
| **Minuto\_aplicado** | **Minuto da partida em que o cartão foi dado** | **INT** | **0 - 100** |
| **Camisa** | **Número da camisa do jogador** | **INT** | **1 - 99 (pode haver ausentes)** |

**Tabela Fato: F\_Estatistica**

|  |  |  |  |
| --- | --- | --- | --- |
| **Coluna** | **Descrição** | **Tipo** | **Domínio** |
| **ID\_Partida** | **Identificador da partida** | **INT** | **-** |
| **ID\_Clube** | **Identificador do clube** | **INT** | **-** |
| **Finalizacao** | **Total de finalizações** | **INT** | **0 - 40** |
| **Finalizacao\_certa** | **Finalizações no alvo** | **INT** | **0 - 30** |
| **Posse\_bola** | **Posse de bola (%)** | **FLOAT** | **0 - 100** |
| **Passes** | **Total de passes** | **INT** | **0 - 1000** |
| **Precisao\_passes** | **Precisão dos passes (%)** | **FLOAT** | **0 - 100** |
| **Faltas\_cometidas** | **Número de faltas cometidas** | **INT** | **0 - 50** |
| **Impedimentos** | **Número de impedimentos** | **INT** | **0 - 10** |
| **Escanteios** | **Número de escanteios** | **INT** | **0 - 15** |

**Tabela Fato: F\_Gols**

|  |  |  |  |
| --- | --- | --- | --- |
| **Coluna** | **Descrição** | **Tipo** | **Domínio** |
| **ID\_Partida** | **Identificador da partida** | **INT** | **-** |
| **ID\_Clube** | **Clube do atleta que marcou o gol** | **INT** | **-** |
| **ID\_Atleta** | **Atleta que marcou o gol** | **INT** | **-** |
| **Minuto\_marcado** | **Minuto do jogo em que o gol foi marcado** | **INT** | **0 - 100** |
| **Tipo\_Gol** | **Tipo de gol** | **STRING** | **Normal, Pênalti, Gol Contra** |

**Tabela Dimensão: D\_Clube**

|  |  |  |  |
| --- | --- | --- | --- |
| **Coluna** | **Descrição** | **Tipo** | **Domínio** |
| **ID\_Clube** | **Identificador do clube** | **INT** | **-** |
| **Nome** | **Nome do clube** | **STRING** | **-** |
| **Estado** | **Estado de origem do clube** | **STRING** | **UF (ex: SP)** |

**Tabela Dimensão: D\_Atleta**

|  |  |  |  |
| --- | --- | --- | --- |
| **Coluna** | **Descrição** | **Tipo** | **Domínio** |
| **ID\_Atleta** | **Identificador do atleta** | **INT** | **-** |
| **Nome** | **Nome do atleta** | **STRING** | **-** |
| **Posição** | **Posição do jogador** | **STRING** | **Goleiro, Zagueiro, Meio-campo, Atacante** |

**Tabela Dimensão: D\_Tecnico**

|  |  |  |  |
| --- | --- | --- | --- |
| **Coluna** | **Descrição** | **Tipo** | **Domínio** |
| **ID\_Tecnico** | **Identificador do técnico** | **INT** | **-** |
| **Nome** | **Nome do técnico** | **STRING** | **-** |