Documentação Técnica - Tabela de Corrida de Taxi Amarelo

Documentação da Tabela: trusted-zone.tb_corrida_taxi_amarelo

Descrição

Tabela que armazena dados das corridas de táxi amarelo processadas a partir de arquivos Parquet brutos localizados no diretório /Volumes/workspace/raw-zone/taxi_amarelo/. Os dados são padronizados, com nomes de colunas traduzidos para português, tipos de dados ajustados e valores numéricos corrigidos para formato decimal com ponto.

A tabela é utilizada para análises de mobilidade urbana, faturamento, padrões de corrida e comportamento de passageiros, integrando informações detalhadas de cada viagem.

Processo de Ingestão e Transformação

- Os arquivos Parquet são carregados do diretório raw.
- Todas as colunas são convertidas para lowercase para padronização.
- Colunas originais são renomeadas conforme o dicionário renomear_colunas para nomes em português e mais descritivos.
- Valores numéricos originalmente com vírgulas são convertidos para pontos para permitir conversão decimal correta.
- Casts seguros são aplicados com base no tipo definido no dicionário tipos_colunas. Valores inválidos para o tipo são convertidos para null.
- Adicionada coluna origem_taxi com valor fixo taxi_amarelo para identificar a origem da base.
- Tabela salva no formato gerenciado do Spark SQL sob o catálogo trusted-zone.

Esquema da Tabela

Nome da Coluna	Tipo de Dados	Descrição
cod_motorista	string	Código identificador do motorista (equivale a vendorid original).
dt_hr_inicio	timestamp	Data e hora do início da corrida (pickup datetime).
dt_hr_fim	timestamp	Data e hora do fim da corrida (dropoff datetime).

Documentação Técnica - Tabela de Corrida de Taxi Amarelo

qtd_pessoas | integer | Quantidade de passageiros na corrida.

dist_percorrida | decimal(10,2) | Distância percorrida durante a corrida em milhas.

cod_taxa | integer | Código do tipo de taxa aplicada na corrida (ratecodeid).

ind_armazenamento | string | Indicador se os dados foram armazenados e encaminhados (store_and_fwd_flag).

cod_bairro_origem | string | Código do bairro ou local de origem da corrida (pulocationid).

cod_bairro_destino | string | Código do bairro ou local de destino da corrida (dolocationid).

tipo_pagamento | string | Forma de pagamento utilizada (ex: cartão, dinheiro).

vlr_taxa_corrida | decimal(10,2) | Valor base da corrida (fare amount).

vlr_taxa_extra | decimal(10,2) | Valor adicional cobrado na corrida (extra).

vlr_taxa_mta | decimal(10,2) | Taxa MTA cobrada (Metropolitan Transportation Authority).

vlr_troco | decimal(10,2) | Valor do troco recebido (tip amount).

vlr_pedagio | decimal(10,2) | Valor pago em pedágios durante a corrida (tolls amount).

cod_taxa_melhoria | integer | Código de taxa adicional de melhoria (improvement surcharge).

vlr_total | decimal(10,2) | Valor total pago na corrida, somando todas as taxas.

vlr_taxa_congestao | decimal(10,2) | Valor da taxa de congestionamento aplicada (congestion surcharge).

vlr_taxa_aeroporto | decimal(10,2) | Valor da taxa cobrada para corridas com destino ou origem em aeroporto.

origem_taxi | string | Indica origem da base, valor fixo: taxi_amarelo.

Observações

- Campos numéricos que não passam no padrão regex para números decimais são convertidos para null para garantir a integridade dos dados.
- Campos de data/hora são convertidos para o tipo timestamp para facilitar consultas temporais e agregações.
- O processo de ingestão sobrescreve a tabela a cada execução para garantir dados atualizados.
- A coluna origem_taxi facilita a união futura com outras tabelas de corridas de táxi, como a de táxi verde.

Uso sugerido

- Consultas temporais para análise de demanda e horário pico.

- Cálculo de receita total por motorista, por bairro, ou por forma de pagamento.
- Análise de padrões de distância e tempo de corrida.
- Monitoramento de taxas extras e suas variações ao longo do tempo.