PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Pós-Graduação em Analytics e Business Inteligence

André Felipe Oliveira Moraes

BOLETIM DE OCORRÊNCIA DE ACIDENTE DE TRÂNSITO COM VITIMA

Belo Horizonte

2023

MÓDULO A – DISCOVERY E PROJETO DE SOLUÇÃO

Contexto do Projeto

Contexto Organizacional: A prefeitura de Belo Horizonte traz em sua base de dados um dataset dos boletins de ocorrência que foram criados durante os acidentes entre veículos automotores que tiveram vítimas em toda cidade de Belo Horizonte.

Motivação: Tais dados podem ser utilizados pela própria prefeitura para melhorar os indicies de acidentes ou por uma companhia de seguros para determinar os lugares onde mais ocorrem acidentes e, posteriormente, fazer uma análise de risco com essas informações no campo das ciências atuárias.

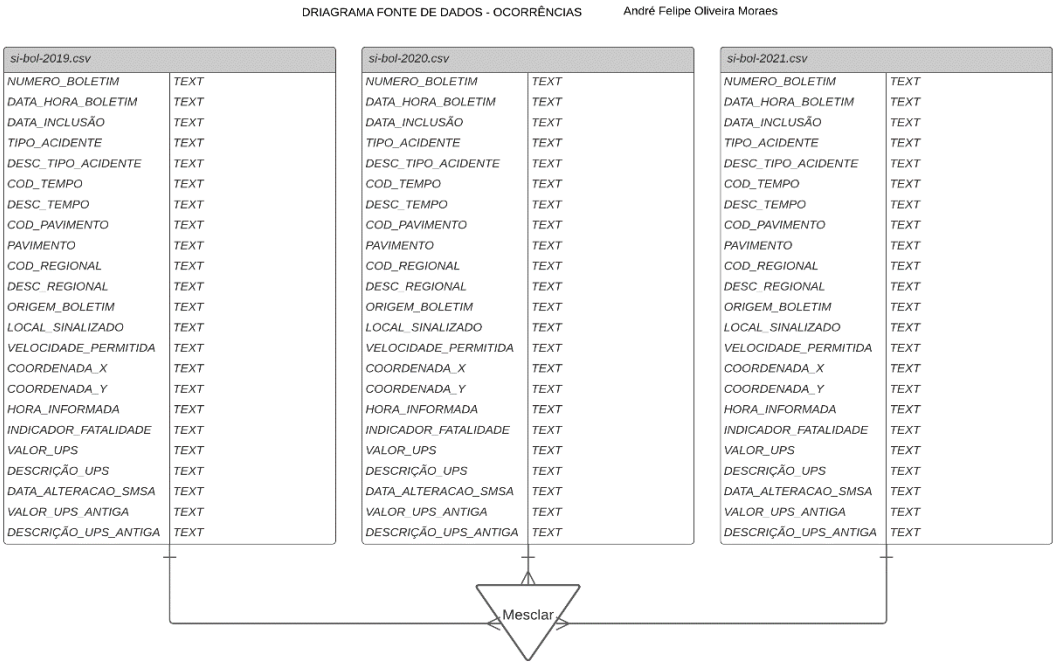
Objetivos estratégicos: Identificar as principais vias que possuem mais acidentes, ocorrências por ano, os bairros com maior número de acidentes, os dias do ano que temos mais acidentes, identificar o tipo de acidente mais comum.

Stakeholders: Coordenadores, gerente e diretores de uma determinada empresa de seguros que possuem clientes em Belo Horizonte.

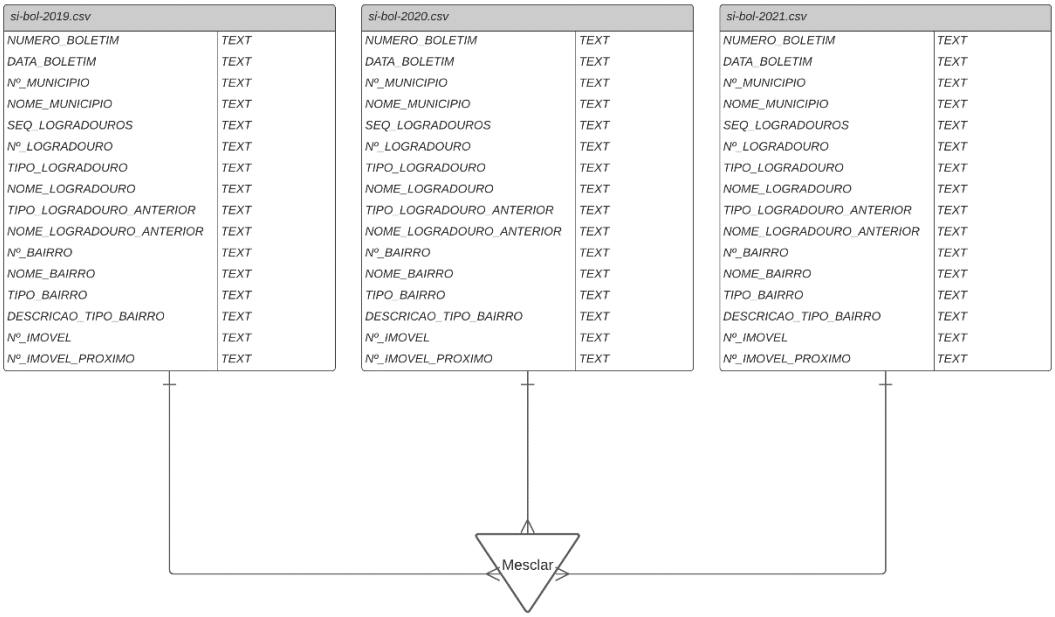
Fonte de dados: A fonte de dados foi toda extraída do site da prefeitura de Belo Horizonte (https://dados.pbh.gov.br/dataset/relacao-de-ocorrencias-de-acidentes-de-transito-com-vitima; https://dados.pbh.gov.br/dataset/relacao-dos-logradouros-dos-locais-de-acidentes-de-transito-com-vitima; https://dados.pbh.gov.br/dataset/relacao-dos-veiculos-envolvidos-nos-acidentes-de-transito-com-vitima) e contempla os anos de 2019, 2020 e 2021.

Modelo de dados

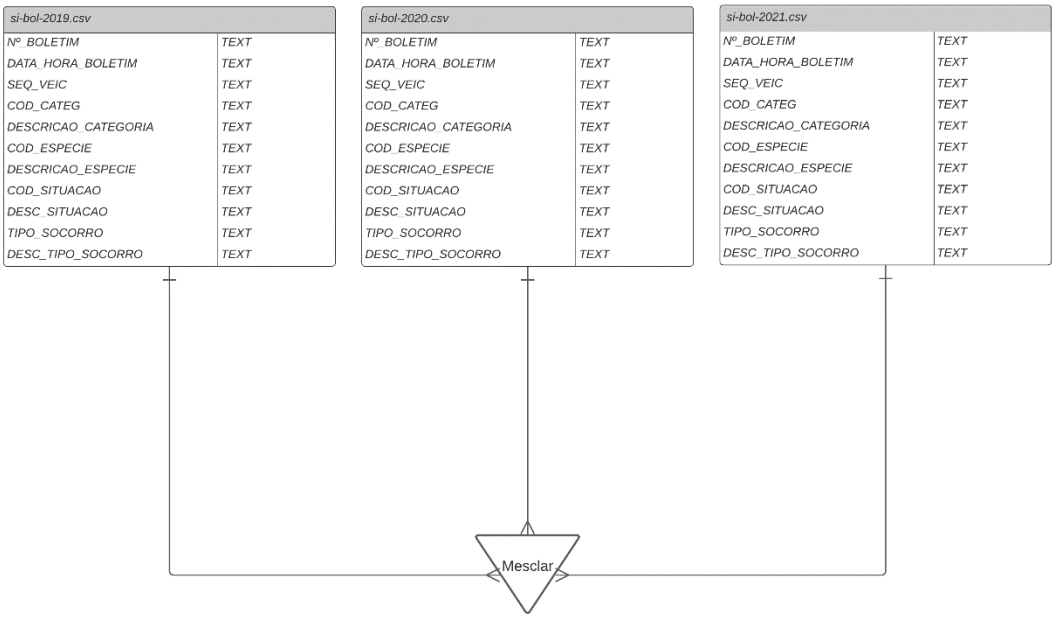
Fonte de Dados:



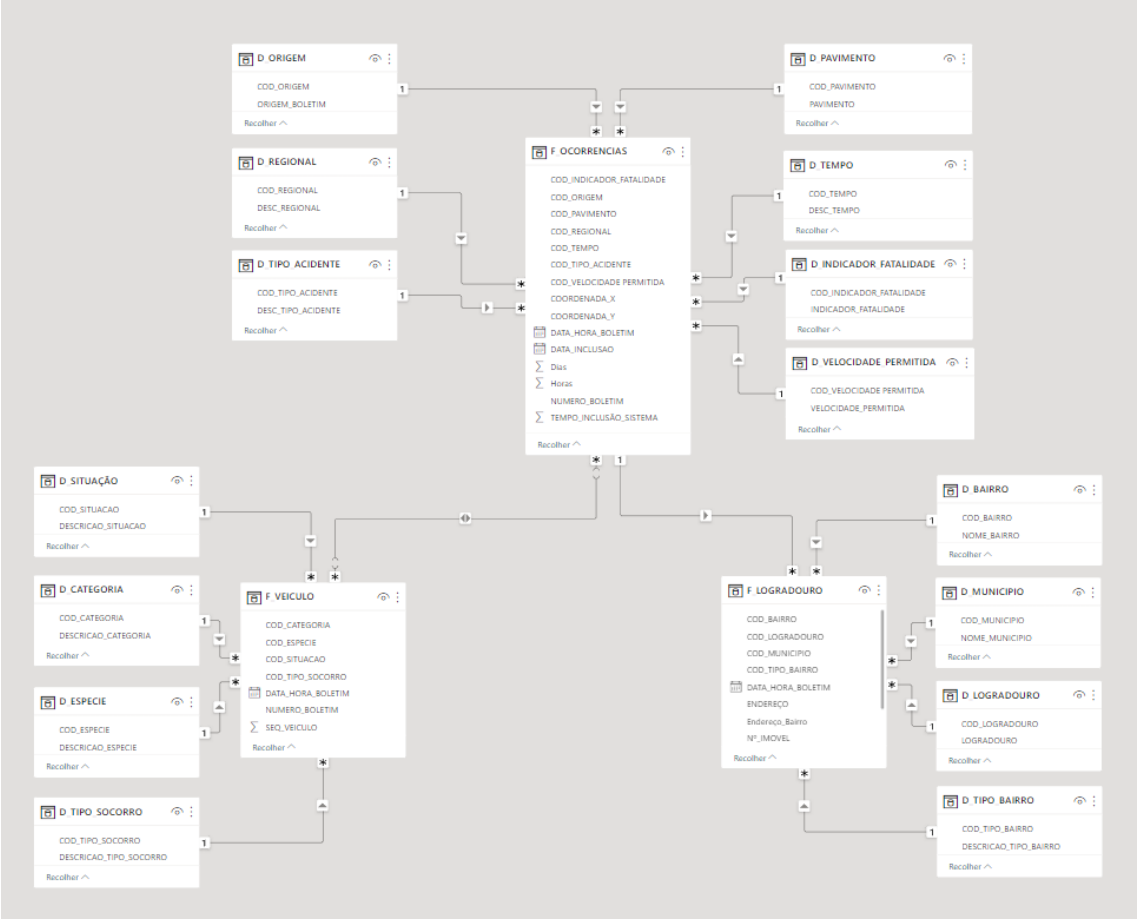
DRIAGRAMA FONTE DE DADOS - LOGRADOURO André Felipe Oliveira Moraes



DRIAGRAMA FONTE DE DADOS - VEICULOS André Felipe Oliveira Moraes



Base Dimensional:



As tabelas fato no projeto são as tabelas:

F_OCORRENCIAS, F_VEICULO e F_LOGRADOURO.

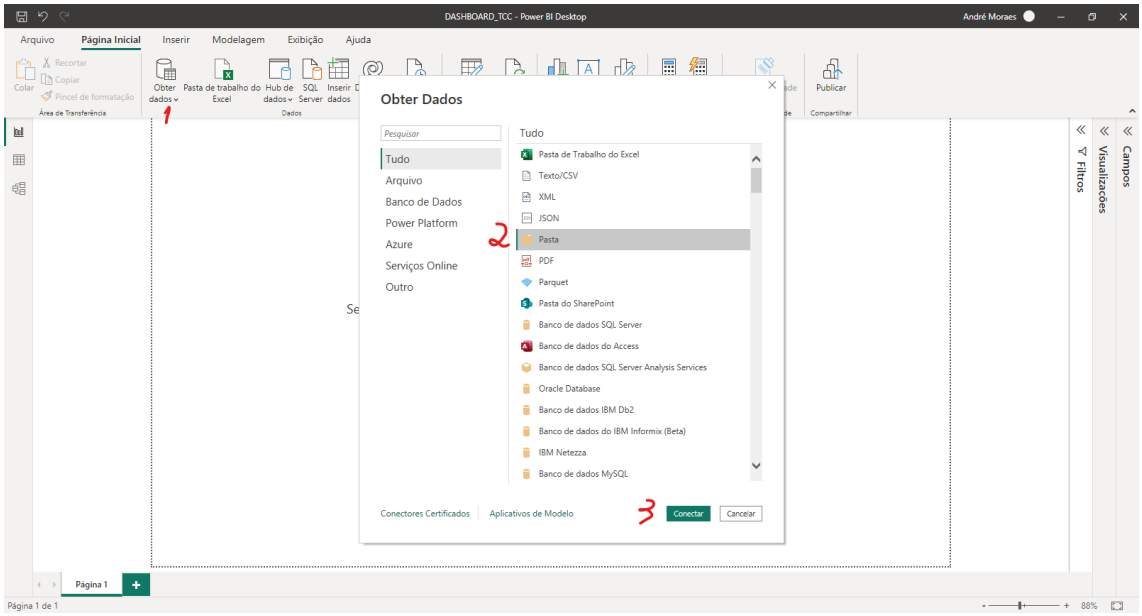
As dimensões são as tabelas:

D_ORIGEM, D_PAVIMENTO, D_REGIONAL, D_TEMPO, D_TIPO_ACIDENTE, D_INDICADOR_FATALIDADE, D_VELOCIDADE_PERMITIDA, D_SITUAÇÃO, D_CATEGORIA, D_ESPECIE, D_TIPO_SOCORRO, D_BAIRRO, D_MUNICIPIO, D_LOGRADOURO e D_TIPO_BAIRRO.

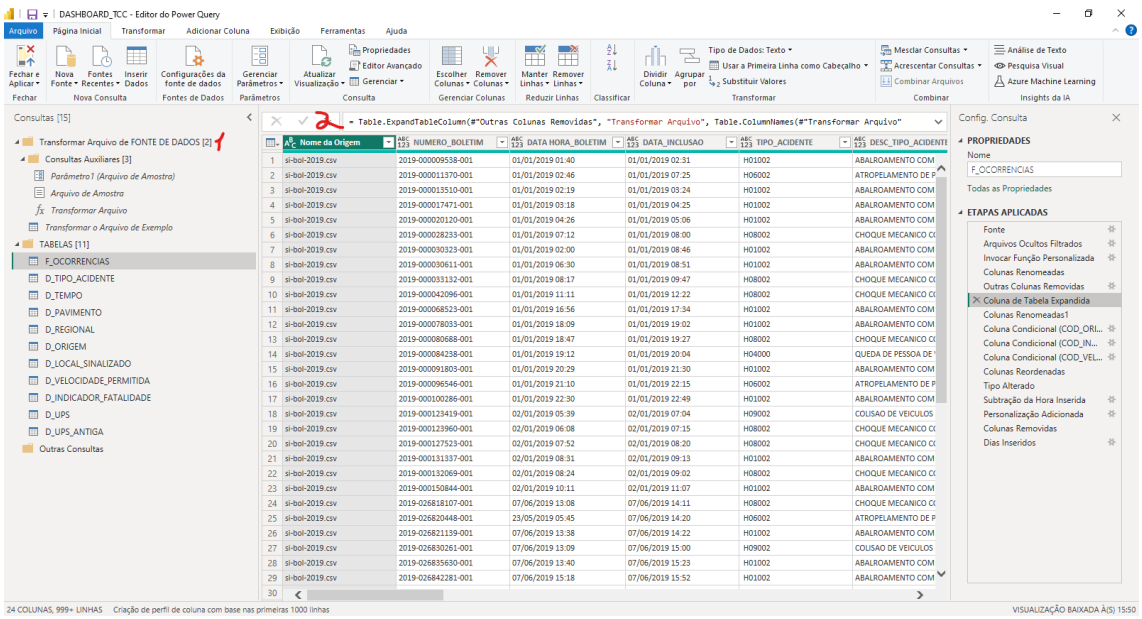
A tabela fato possui a sigla **F** no começo do nome e as dimensões a sigla **D** no começo do nome para facilitar a sua identificação, sendo que o modelo dimensional é o Star Schema ou esquema de estrela.

Processos de Integração, Tratamento e Carga de Dados

Ingestão de Dados e Processos ETL – A ingestão de dados e os processos ETL ocorreram diretamente na ferramenta Microsoft Power BI no qual foi possível o tratamento e carga dos dados utilizando o Power Query. A ingestão de dados foi feita utilizando a opção de pasta, do qual faz a leitura dos arquivos que estão na pasta na qual estão salvas a fonte de dados que são os arquivos sl-bol-2019.csv, sl-bol-2020.csv e sl-bol-2021.csv, conforme a seguir:



Com isso, o Power BI gerou uma transformação de arquivos que conseguiram ler os arquivos da pasta no formato de csv e os agrupou cada base de dados em uma única tabela que foram denominadas como **F_OCORRENCIAS, F_VEICULO e F_LOGRADOURO** que gerou a coluna “Nome da Origem” que é referente ao arquivo da origem e a cada linha presente no arquivo, conforme a seguir:



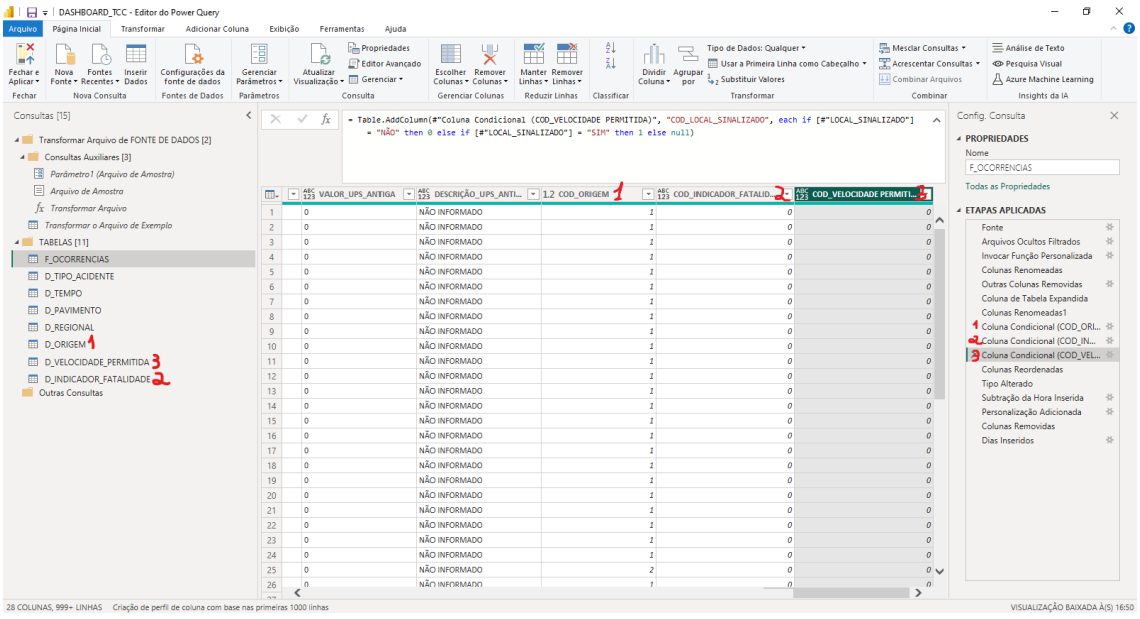
ETL F_OCORRÊNCIAS:

Em seguida, foram renomeadas algumas colunas que apresentavam um espaçamento em branco na frente do nome e que poderiam atrapalhar no desenvolvimento de medidas e análises ao decorrer do trabalho. Em suma, todas as colunas tiveram seu nome alterado para retirar esse espaço, com exceção da coluna Nome da Origem que não tinha esse problema e da coluna " TIPO_ACIDENTE", que passou a ter no seu inicio o sufixo “COD_” para manter a padronização da base de dados.

Foi adicionado uma coluna condicional denominada como “COD_ORIGEM”, tal coluna foi criada para atribuir ID a coluna “ORIGEM_BOLETIM” atribuindo os códigos 0, 1 e 2 para cada situação, sendo o 1 para “POLÍCIA MILITAR”, 2 para “POLÍCIA CIVIL” e 0 para “NI”, ou seja, não informado. Com isso, foi possível criar a dimensão D_ORIGEM.

Também, foi adicionado outra coluna condicional denominada como “COD_INDICADOR_FATALIDADE”, do qual atribuiu ID a coluna “INDICADOR_FATALIDADE” com os códigos 0 e 1, sendo o 0 para “NÃO” e o 1 para “SIM”. Isso possibilitou a criação da dimensão D_INDICADOR_FATALIDADE.

Por fim, foi adicionado uma nova coluna condicional nomeada como “COD_VELOCIDADE_PERMITIDA”, da qual criou ID para a coluna “VELOCIDADE_PERMITIDA” que atribuiu os códigos 0,1,2,3,4,5,6,7 e 8, sendo o 0 de “0”, 1 de “20”, 2 de “30”, 3 de “40”, 4 de “50”, 5 de “60”, 6 de “70”, 7 de “80” e 8 de “110”. Portanto, essas seriam as medidas dos agentes de trânsito da velocidade da via, porém, cabe ressaltar que não existe via com velocidade permitida igual a 0 KM/H e, portanto, o dado pode ser tido como não informado pelo agente. Cabe ressaltar que com essa coluna foi possível criar a dimensão D_VELOCIDADE_PERMITIDA.



Em seguida, as colunas da tabela F_OCORRENCIAS foram reordenadas para que as colunas condicionais que foram acrescentadas na base ficassem perto da coluna correspondente para facilitar as etapas seguintes.

Foram alterados os tipos dos dados, visto que todos estavam como genérico que é o “ABC123” que o Power BI atribui automaticamente e passaram a ser do tipo Texto para todos que são Ids como “NUMERO_BOLETIM” e as colunas que começavam como “COD_”,

além de alterar para o tipo DateTime os campos “DATA_HORA_BOLETIM” e “DATA_INCLUSAO”.

Em seguida, foi criado uma nova coluna denominada como TEMPO_INCLUSAO_SISTEMA do qual fazia a subtração da DATA_INCLUSAO com a DATA_HORA_BOLETIM e gerava o tempo de duração entre as duas datas.

Também, houve um tratamento na duração para caso a mesma apresentação tempo negativo, ou seja, menor que 0 e, portanto, ele passou a representar 0, visto que se a ocorrência foi incluída antes do seu acontecimento se trata de um erro de preenchimento.

Na etapa seguinte as colunas de Descrição das dimensões foram excluídas, além da coluna “Nome da Origem” que não era mais necessário, visto que os dados já estavam mesclados entre si e o número da ocorrência já conta com o ano do dado no seu início e as colunas VALOR_UPS e VALOR_UPS_ANTIGA que apresentavam valores igual a 0, sendo assim, sem nenhuma atribuição prática para o projeto. O campo “DATA_ALTERACAO_SMSA” que estava com o valor de “00/00/0000” para todos os registros e, portanto, presume-se que seja um dado para saber quando a base de dados passou por uma alteração e por não haver registros válidos entende-se que não sofreu alteração, logo é um dado irrelevante para a análise.

Por fim, foi acrescentado a coluna de Dias que é justamente o número de dias que cada ocorrência demorou para entrar no sistema.

ETL D_TIPO_ACIDENTE, D_TEMPO, D_PAVIMENTO e D_REGIONAL,

Essas tabelas tiveram um ETL praticamente igual, com isso resolvi reuni-los aqui para explicar os pontos em comum entre eles. Essas dimensões foram criadas a partir da duplicação da tabela fato F_OCORRENCIAS, com isso as etapas iniciais são as mesmas da tabela de origem.

Tal opção foi feita para que os dados não se perdessem na hora de excluir as colunas de descrição da tabela fato, visto que o Power BI tem a opção de “Duplicar” e “Referenciar”, sendo que em “Duplicar” ele faz a duplicação da tabela selecionada com todas as etapas que ela sofreu e na opção de “Referenciar” ela apenas utiliza a tabela depois de todo o tratamento sem as etapas de tratamento que ela sofreu.

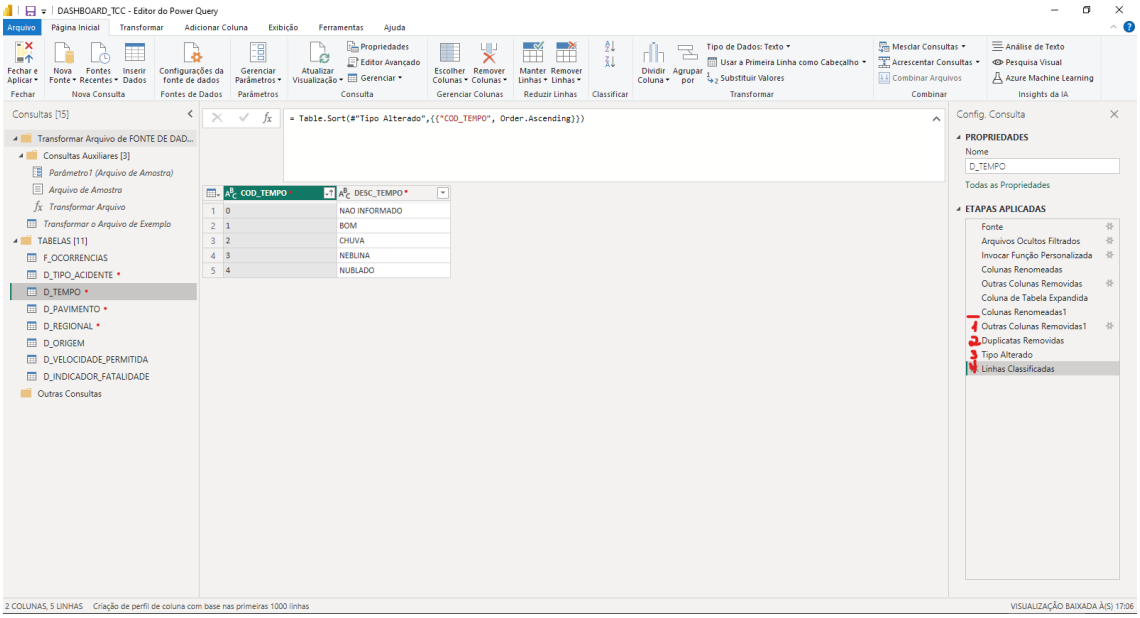
Portanto, até a etapa “Colunas Renomeadas1” as dimensões tem o mesmo ETL que a tabela fato, ou seja, temos todo o trabalho da leitura dos arquivos da pasta e a renomeação para adequação das informações. Em seguida, são excluídas todas as colunas que não fazem parte da dimensão e são mantidos a coluna de ID que tem como inicio “COD_” e a sua respectiva descrição.

Por conseguinte, temos a remoção das duplicatas dos registros o que mantem apenas o ID e a descrição distinto da base inteira e possibilidade fazer a conexão dimensional de um para muitos.

Em seguida, as colunas tem os seus tipos alterados de genérico “ABC123” para o tipo texto “ABC”, apesar dos Ids contarem com números em sua maioria, estes não sofrerão qualquer tipo de operação matemática e, por isso, passam a serem texto.

Na tabela D_REGIONAL tem um passo a mais que as outras tabelas, pois o valor que estava em branco foi substituído por “NI”, ou seja, não informado.

Por fim, foi feito a classificação das linhas em ordem crescente para facilitar a identificação dos registros e de seus Ids na hora da leitura e manuseio dos dados.



ELT D_ORIGEM, D_LOCAL_SINALIZADO, D_VELOCIDADE_PERMITIDA E D_INDICADOR_FATALIDADE

Assim como os ETLs que foram feitos nas outras dimensões, essas tabelas foram criadas a partir da duplicação da tabela fato F_OCORRENCIAS, com isso as etapas iniciais são as mesmas da tabela de origem.

Tal opção foi feita para que os dados não se perdessem na hora de excluir as colunas de descrição da tabela fato, visto que o Power BI tem a opção de “Duplicar” e “Referenciar”, sendo que em “Duplicar” ele faz a duplicação da tabela selecionada com todas as etapas que ela sofreu e na opção de “Referenciar” ela apenas utiliza a tabela depois de todo o tratamento sem as etapas de tratamento que ela sofreu.

Nessas tabelas todas possuem a etapa de criação de coluna condicional, visto que na base de dados original não havia as colunas de Id para esses assuntos, portanto, todos os ETLs dessas tabelas possuem esse passo a mais que é apenas para a criação da coluna de ID.

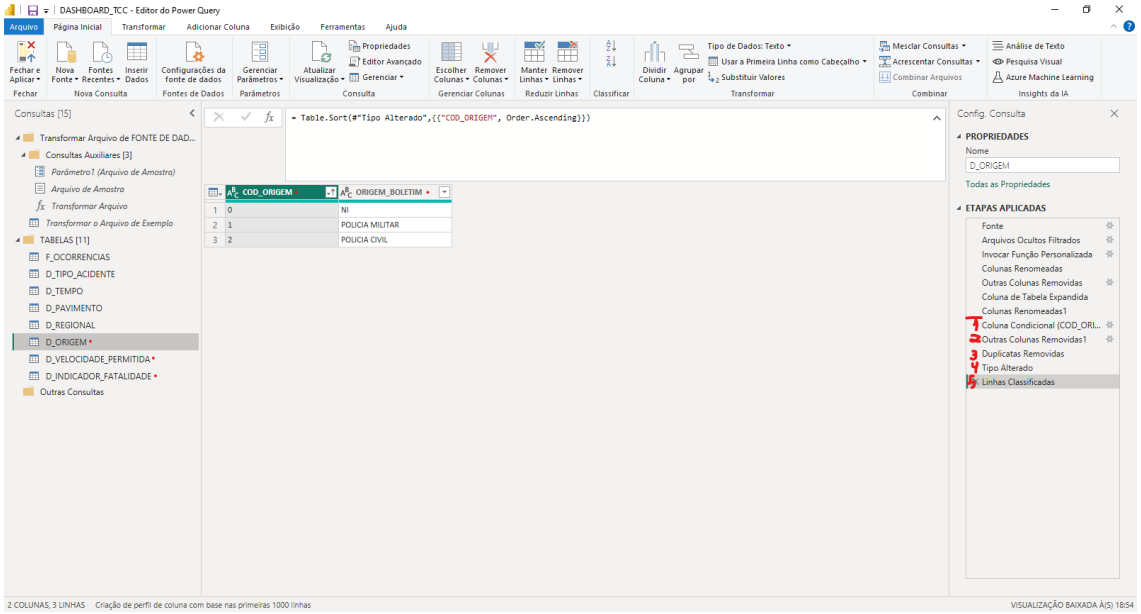
Portanto, até a etapa “Colunas Renomeadas1” as dimensões tem o mesmo ETL que a tabela fato, ou seja, temos todo o trabalho da leitura dos arquivos da pasta e a renomeação para adequação das informações. Em seguida, são excluídas todas as colunas que não fazem parte da dimensão e são mantidos a coluna de ID que tem como início “COD_” e a sua respectiva descrição.

Por conseguinte, temos a remoção das duplicatas dos registros o que mantem apenas o ID e a descrição distinto da base inteira e possibilidade fazer a conexão dimensional de um para muitos.

Em seguida, as colunas tem os seus tipos alterados de genérico “ABC123” para o tipo texto “ABC”, apesar dos Ids contarem com números em sua maioria, estes não sofrerão qualquer tipo de operação matemática e, por isso, passam a serem texto.

Na tabela D_REGIONAL tem um passo a mais que as outras tabelas, pois o valor que estava em branco foi substituído por “NI”, ou seja, não informado.

Por fim, foi feito a classificação das linhas em ordem crescente para facilitar a identificação dos registros e de seus Ids na hora da leitura e manuseio dos dados.

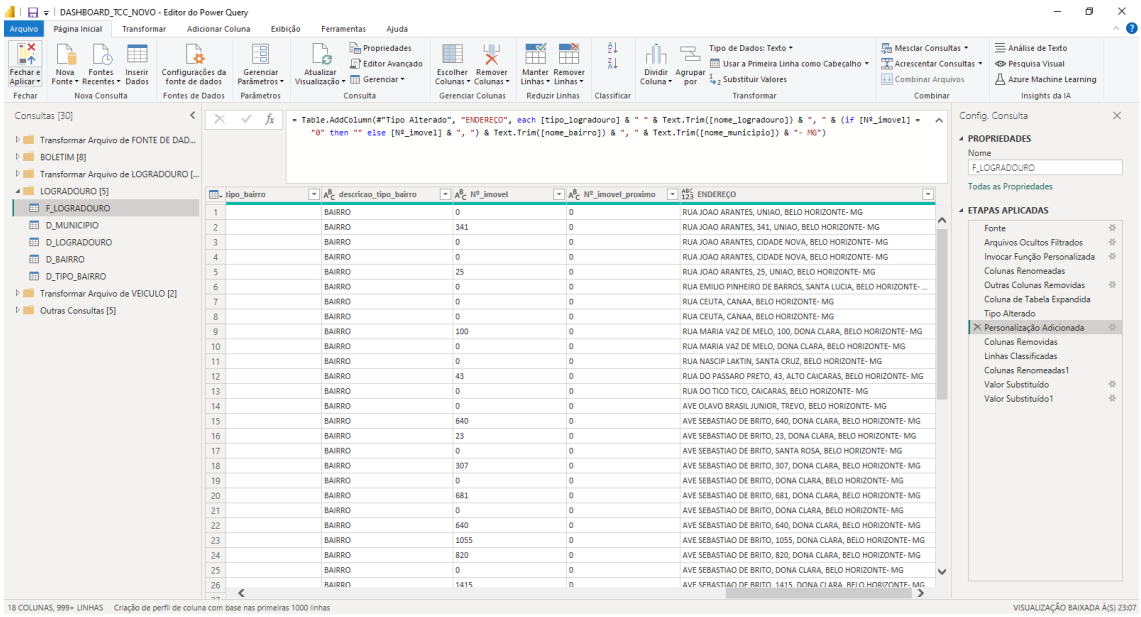


ETL F_LOGRADOURO:

Assim como F_OCORRENCIAS, a tabela fato F_LOGRADOURO passou pelas mesmas transformações para a leitura da base de dados na pasta de logradouro.

Em seguida, o tipo de todas as colunas foi alterado, visto que todos estavam como genérico que é o “ABC123” que o Power BI atribui automaticamente e passaram a ser do tipo Texto para todos com exceção para data_boletim que passou a ser datetime e seq_logradouros que passou a ser do tipo int.

Foi adicionado a coluna denominada como “ENDEREÇO”, tal coluna foi criada para formar o endereço dos locais de acidente combinando os dados do logradouro e removendo o espaçamento em branco nas células.



As colunas de descrição das dimensões foram excluídas, bem como as colunas Nome da Origem, tipo_logradouro_anterior, nome_logradouro_anterior, nº_imovel_proximo e tipo_logradouro, pois estes não serão usados na análise.

Em seguida, as linhas foram classificadas de forma ascendente pelo Nº_boletim.

As colunas foram renomeadas para seguir o padrão da tabela F_OCORRENCIAS, tais como acréscimo de COD_, bem como colocar as colunas em maiúsculo e padronizar as colunas número boletim e data boletim.

Por fim, o endereço que começava com PCA SETE DE SETEMBRO foi substituído por AVE AFONSO PENA, visto que não existe um logradouro com o endereço do primeiro e, portanto, foi necessário fazer a correção para que o mapa lesse corretamente.

ELT D_MUNICIPIO, D_LOGRADOURO, D_BAIRRO E D_TIPO_BAIRRO

Essas dimensões foram criadas a partir da duplicação da tabela fato F_LOGRADOURO, com isso as etapas iniciais são as mesmas da tabela de origem.

Tal opção foi feita para que os dados não se perdessem na hora de excluir as colunas de descrição da tabela fato, visto que o Power BI tem a opção de “Duplicar” e “Referenciar”, sendo que em “Duplicar” ele faz a duplicação da tabela selecionada com todas as etapas que ela sofreu e na opção de “Referenciar” ela apenas utiliza a tabela depois de todo o tratamento sem as etapas de tratamento que ela sofreu.

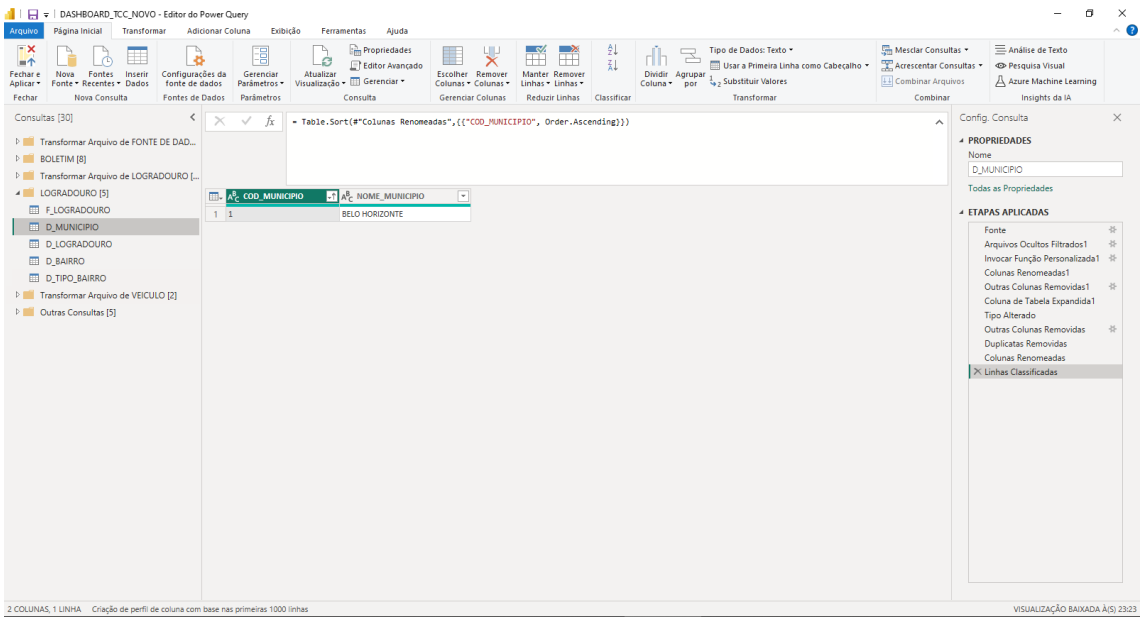
Portanto, até a etapa “Tipo Alterado” as dimensões tem o mesmo ETL que a tabela fato, ou seja, temos todo o trabalho da leitura dos arquivos da pasta e a renomeação para adequação das informações. Em seguida, são excluídas todas as colunas que não fazem parte da dimensão e são mantidos a coluna de ID e a sua respectiva descrição.

Por conseguinte, temos a remoção das duplicatas dos registros o que mantem apenas o ID e a descrição distinto da base inteira e possibilidade fazer a conexão dimensional de um para muitos.

Em seguida, as colunas são renomeadas para ficarem maiúsculo e adequarem ao ID da tabela fato.

Na tabela D_LOGRADOURO tem um passo a mais que as outras tabelas, pois as colunas tipo_logradouro e nome_logradouro são combinados pra formarem o logradouro.

Por fim, foi feito a classificação das linhas em ordem crescente para facilitar a identificação dos registros e de seus Ids na hora da leitura e manuseio dos dados.



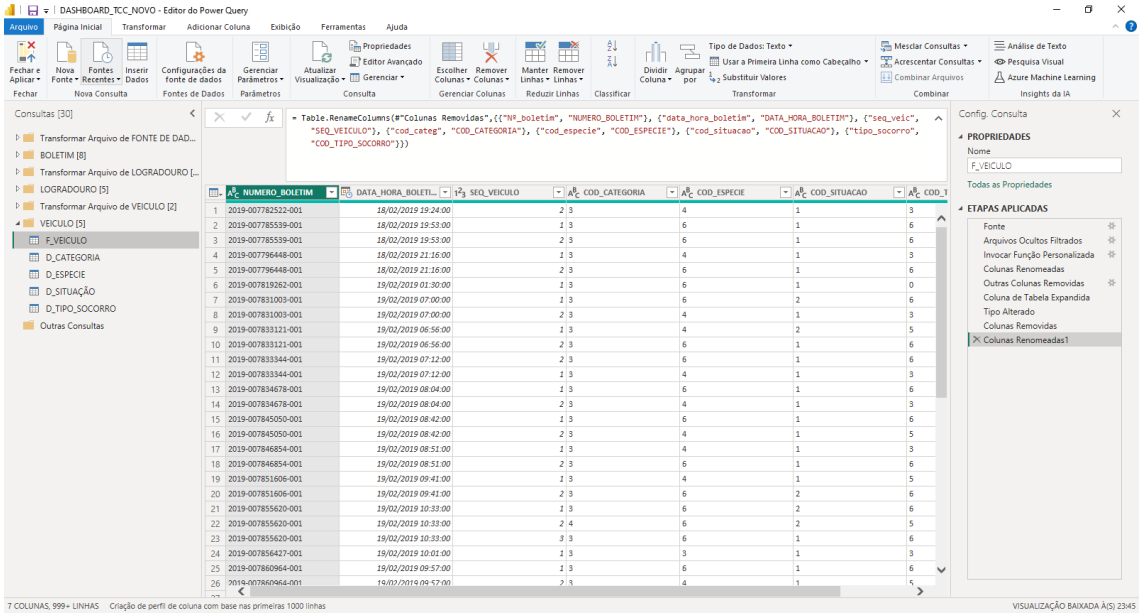
ETL F_VEICULO:

Assim como F_OCORRENCIAS, a tabela fato F_VEICULO passou pelas mesmas transformações para a leitura da base de dados na pasta de logradouro.

Em seguida, o tipo de todas as colunas foi alterado, visto que todos estavam como genérico que é o “ABC123” que o Power BI atribui automaticamente e passaram a ser do tipo Texto para todos com exceção para data_hora_boletim que passou a ser datetime e seq_veic que passou a ser do tipo int.

As colunas de descrição das dimensões foram excluídas, bem como a coluna Nome da Origem, pois esta não será usada na análise.

Por fim, as colunas foram renomeadas para seguir o padrão da tabela F_OCORRENCIAS, tais como acréscimo de COD_, bem como colocar as colunas em maiúsculo e padronizar as colunas número boletim e data boletim.



ELT D_CATEGORIA, D_ESPECIE, D_SITUACAO E D_TIPO_SOCORRO

Essas dimensões foram criadas a partir da duplicação da tabela fato F_VEICULO, com isso as etapas iniciais são as mesmas da tabela de origem.

Tal opção foi feita para que os dados não se perdessem na hora de excluir as colunas de descrição da tabela fato, visto que o Power BI tem a opção de “Duplicar” e “Referenciar”, sendo que em “Duplicar” ele faz a duplicação da tabela selecionada com todas as etapas que ela sofreu e na opção de “Referenciar” ela apenas utiliza a tabela depois de todo o tratamento sem as etapas de tratamento que ela sofreu.

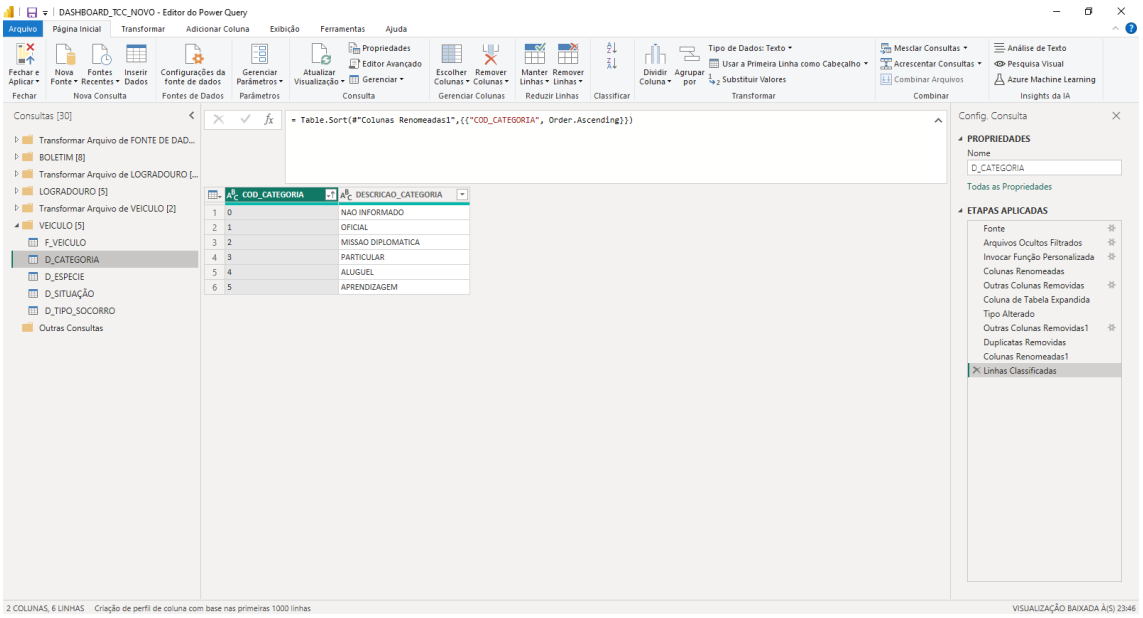
Portanto, até a etapa “Tipo Alterado” as dimensões tem o mesmo ETL que a tabela fato, ou seja, temos todo o trabalho da leitura dos arquivos da pasta e a renomeação para adequação das informações. Em seguida, são excluídas todas as colunas que não fazem parte da dimensão e são mantidos a coluna de ID e a sua respectiva descrição.

Por conseguinte, temos a remoção das duplicatas dos registros o que mantem apenas o ID e a descrição distinto da base inteira e possibilidade fazer a conexão dimensional de um para muitos.

Em seguida, as colunas são renomeadas para ficarem maiúsculo e adequarem ao ID da tabela fato.

Na tabela D_SITUAÇÃO tem um passo a mais que as outras tabelas, pois os valores do campo de DESCRICAO_SITUACAO são substituídos no caso de ESTACIONADO e vazio para apenas “ESTACIONADO”.

Por fim, foi feito a classificação das linhas em ordem crescente para facilitar a identificação dos registros e de seus Ids na hora da leitura e manuseio dos dados.



Códigos Fonte (Link para repositório externo)

Repositório da Prefeitura de Belo Horizonte com os boletins de ocorrência:

https://dados.pbh.gov.br/dataset/relacao-de-ocorrencias-de-acidentes-de-transito-com-vitima

Repositório da Prefeitura de Belo Horizonte com os logradouros dos boletins de ocorrência:

https://dados.pbh.gov.br/dataset/relacao-dos-logradouros-dos-locais-de-acidentes-de-transito-com-vitima

Repositório da Prefeitura de Belo Horizonte com os veículos dos boletins de ocorrência:

https://dados.pbh.gov.br/dataset/relacao-dos-veiculos-envolvidos-nos-acidentes-de-transito-com-vitima

Dashboard com o tratamento de dados no Google Drive:

https://drive.google.com/drive/folders/1AFe6HQDRMgajhu6v_ul3YGLTu7FLlkIi?usp=sharing

Fonte de dados no Google Drive:

https://drive.google.com/drive/folders/19bc7xSXAjYxTzn60nErgN8JBDjBo3O8X?usp=sharing