



DOCUMENTAÇÃO DO MVP
Sprint: Engenharia de Dados

DOCUMENTAÇÃO DO MVP

Sprint: Engenharia de Dados

Aluno: Rafael Ferraz de Queiroz

Brasília, outubro de 2023

Sumário

Lista de ilustrações.....	3
Lista de Tabelas	4
1. Objetivo.....	5
2. Problema	5
3. Bases de Dados	6
4. Ferramentas	6
5. Coleta de Dados	7
6. Modelagem	8
6.1. Modelagem no BigQuery	8
6.2. Modelagem no Cloud SQL.....	10
7. Catálogo de Dados	14
8. Carga	18
8.1. Etapas da Carga.....	19
8.1.1. Carga das Dimensões	19
8.1.2. Carga da Fato	23
9. Plano de Execução e Agendamento.....	27
10. Análise.....	29
10.1. Qualidade dos Dados	29
10.2. Solução do Problema	37
10.2.1. Perguntas	37
10.2.2. Discussão Geral Sobre as Perguntas	46
11. Auto avaliação.....	47

Lista de ilustrações

Figura 1- Interface do Google Storage com os datasets armazenados.	7
Figura 2- Evidência da execução dos comandos.....	9
Figura 3- Detalhamento da tabela de “ocorrencias”.....	9
Figura 4- Detalhamento da tabela de “ocorrencias”.....	10
Figura 5- Instância ativa do MySQL no Cloud SQL.....	10
Figura 6 - Configuração da instância.....	11
Figura 7- Diagrama do banco "datatran" desenhado com a ferramenta GenMyModel.....	12
Figura 8- Modelo de dados carregado no DBeaver.....	13
Figura 9- Interface da página inicial do Dataprep.....	18
Figura 10- Em destaque, as etapas da carga das dimensões.....	19
Figura 11- Leitura dos dados.....	20
Figura 12 - Interface de transformação dos dados da tabela dm_uf.....	21
Figura 13 - Interface de conexão de saída dos dados.....	21
Figura 14- Tela inicial de configuração do Output.....	22
Figura 15- Opção para incluir um script SQL antes ou após a carga.....	22
Figura 16- Consulta na tabela dm_uf pelo DBeaver.....	23
Figura 17- Fluxo de carga da tabela Fato.....	24
Figura 18- Configuração do recipe “carga_fato”, com as transformações aplicadas.....	25
Figura 19 - Comportamento da opção "Lookup" no Dataprep.....	26
Figura 20- Consulta à tabela de ft_ocorrencias no DBeaver.....	27
Figura 21- Planos de execução no módulo Plans.....	27
Figura 22- Sequência de execução. (De cima para baixo).....	28
Figura 23 -Interface de agendamento de execuções do módulo Dataprep.....	28
Figura 24 - Histórico de execuções do Dataprep.....	29

Lista de Tabelas

Tabela 1 - dm_causaacidente	14
Tabela 2 - dm_municipio	14
Tabela 3 - dm_tipoacidente	15
Tabela 4 - dm_tracadovia	15
Tabela 5 - dm_uf	15
Tabela 6 - ft_ocorrencias	18
Tabela 7 - Análise de qualidade da Tabela dm_causaacidente	30
Tabela 8 - Análise de qualidade da Tabela dm_municipio	30
Tabela 9 - Análise de qualidade da Tabela dm_tipoacidente	31
Tabela 10 - Análise de qualidade da Tabela dm_tracadovia	31
Tabela 11 - Análise de qualidade da Tabela dm_uf	32
Tabela 12 - Pergunta 1	37
Tabela 13 - Pergunta 2	40
Tabela 14 - Pergunta 3	40
Tabela 15 - Pergunta 4	41
Tabela 16 - Pergunta 5	41
Tabela 17 - Pergunta 6	42
Tabela 18 - Pergunta 7	43
Tabela 19 - Pergunta 8	44
Tabela 20 - Pergunta 9	45
Tabela 21 - Pergunta 10	46

1. Objetivo

Este trabalho tem o intuito de apresentar de forma detalhada todo o processo para o desenvolvimento de um pipeline de dados em plataforma de nuvem, contemplando a busca dos dados, transformação, análise de qualidade, resolução do problema e por fim, uma discussão sobre os resultados alcançados acerca dos problemas que foram levantados.

2. Problema

Para elaboração das perguntas, escolhi analisar o histórico de acidentes de trânsito de veículos ocorridos nas estradas brasileiras. Abaixo seguem os 10 questionamentos e uma breve descrição sobre cada um deles:

1. Qual o evolutivo de acidentes por ano?

Agrupar a quantidade de acidentes ocorridos por ano, considerando os últimos 10 anos ou anos disponíveis nas bases de dados da PRF.

2. Qual o evolutivo de vítimas em acidentes por ano e mês, considerando o quantitativo de pessoas envolvidas X fatalidades?

Comparativo cronológico por ano/mês sobre a quantidade de pessoas envolvidas em acidentes, paralelo ao quantitativo de pessoas que vieram a óbito, considerando os últimos 5 anos ou anos disponíveis na base da PRF.

3. Qual o percentual de pessoas envolvidas em acidentes por gênero?

Percentual do total de vítimas em acidentes de toda a base coletada por gênero masculino, feminino e outros (sem informação de gênero na base).

4. Quais são as “Top 10” causas de acidente?

10 maiores causas de acidentes considerando toda a base coletada.

5. Quais são os “Top 20” modelos de veículos envolvidos em acidentes?

20 veículos mais envolvidos em acidentes considerando toda a base coletada.

6. Qual o quantitativo de acidentes por dia da semana?

Quantitativo de acidentes por dia da semana (Domingo, Segunda-feira, (...), Sexta-feira) considerando toda a base coletada.

7. Qual o quantitativo de acidentes por faixa de horário?

Comparativo de acidentes ocorridos por faixa de horário (Exemplo: 13 = 13:00 até 13:59), considerando toda a base coletada.

8. Qual o quantitativo de acidentes por faixa etária?

Volumetria de vítimas em acidentes por faixa etária (Exemplo: 0 a 5, 6 a 10, 11 a 15, 16 a 20, 21 a 25, (...), 80 e mais)

9. Qual o quantitativo de vítimas X Fatalidades em acidentes por Estado?

Comparativo do quantitativo de vítimas e vítimas fatais por Estado.

10. Quais são as 10 maiores causas de fatalidades nos acidentes?

As 10 maiores causas de acidentes considerando a quantidade de vítimas que vieram a óbito.

Estes questionamentos foram preservados no decorrer do desenvolvimento da solução para fomentar a discussão acerca dos resultados obtidos.

3. Bases de Dados

Para elucidação dos questionamentos propostos no problema, optei pela utilização dos dados do Boletim de Acidentes de Trânsito (BAT) disponibilizados pela Polícia Rodoviária Federal através do portal de Dados Abertos da PRF: <https://www.gov.br/prf/pt-br/aceso-a-informacao/dados-abertos/dados-abertos-da-prf>.

Segundo a Polícia Rodoviária Federal – PRF, O registro de acidentes é realizado através do sistema BAT, que coleta informações referentes aos envolvidos (identificação, estado físico, se era passageiro, condutor, etc.), ao local, aos veículos, à dinâmica do acidente, etc. Os dados disponíveis têm origem nos sistemas BR-Brasil e BAT. O sistema BR-Brasil foi utilizado em nível nacional entre 2007 e 2016. O sistema BAT é utilizado desde 2017.

Fonte: <https://www.gov.br/prf/pt-br/aceso-a-informacao/dados-abertos/dicionario-acidentes>

4. Ferramentas

Para o desenvolvimento da solução, foi escolhida a suíte de computação em nuvem [Google Cloud Plataform](#). Entre os motivos para a escolha da suíte estão os créditos de avaliação gratuita no valor de US\$ 300 (equivalente a R\$ 1.669,00 na data da inscrição) e o prazo de 3 meses para utilização destes créditos, além da curiosidade em explorar os recursos disponibilizados pela plataforma.

As seguintes ferramentas da plataforma foram utilizadas para o gerenciamento dos dados:

- Cloud Storage: Para armazenamento dos datasets em formato CSV obtidos do portal da PRF.
- BigQuery: Para armazenamento centralizado da série histórica de dados.
- Datapret: Ferramenta de ETL para a preparação dos dados.
- Cloud SQL: Serviço de banco de dados relacional com a estrutura de Data warehouse.
- Cloud Shell: Shell bash on-line baseado em Debian.

5. Coleta de Dados

A coleta dos dados foi realizada de forma manual, acessando o Portal de Dados Abertos da PRF (<https://www.gov.br/prf/pt-br/aceso-a-informacao/dados-abertos/dados-abertos-da-prf>) e baixando as planilhas de “Documento CSV de Acidentes 2023 (Agrupados por ocorrência)” dos últimos 4 anos (2020 até 2023).

Em seguida os dados foram armazenados manualmente no serviço de armazenamento *Google Cloud Storage* em um Bucket chamado “rafaelferraz-storage”, em um diretório chamado “prf”, conforme evidência abaixo:

rafaelferraz-storage

Local	Classe de armazenamento	Acesso público	Proteção
us-east1 (Carolina do Sul)	Standard	Não público	Nenhum

OBJETOS CONFIGURAÇÃO PERMISSÕES PROTEÇÃO CICLO DE VIDA OBSERVABILIDADE RELATÓRIOS DE INVENTÁRIO

Intervalos > rafaelferraz-storage > prf

FAZER UPLOAD DE ARQUIVOS CARREGAR PASTA CRIAR PASTA TRANSFERIR DADOS GERENCIAR RETENÇÕES FAZER O DOWNLOAD EXCLUIR

Filtrar apenas pelo prefixo do nome Filtro Filtrar objetos e pastas Mostrar dados excluídos













<input type="checkbox"/>	Nome	Tamanho	Tipo	Criado	Classe de armazenamento	Última modificação	A
<input type="checkbox"/>	 datatran2020.csv	17,5 MB	text/csv	29 de set. de 2023 22:20:47	Standard	29 de set. de 2023 22:20:47	N  
<input type="checkbox"/>	 datatran2021.csv	17,8 MB	text/csv	29 de set. de 2023 22:20:49	Standard	29 de set. de 2023 22:20:49	N  
<input type="checkbox"/>	 datatran2022.csv	17,8 MB	text/csv	29 de set. de 2023 22:20:49	Standard	29 de set. de 2023 22:20:49	N  
<input type="checkbox"/>	 datatran2023.csv	12,2 MB	text/csv	29 de set. de 2023 22:21:42	Standard	29 de set. de 2023 22:21:42	N  

Figura 1- Interface do Google Storage com os datasets armazenados.

6. Modelagem

Foram criados dois esquemas para armazenamento dos arquivos CSV em duas diferentes ferramentas do Google Cloud, sendo uma tabela *Flat* no *BigQuery* e um modelo de dados do tipo “Estrela” no serviço de banco de dados relacional *Cloud SQL*.

6.1. Modelagem no BigQuery

Optei por unir os dados dos arquivos CSV em uma tabela Flat na ferramenta Google BigQuery denominada “ocorrencias”, centralizando-os em sua forma bruta (sem tratamento de tipos e de valores), em um conjunto de dados denominado “datatran_db”.

Além disso, a tabela foi particionada por data, levando em consideração a data da ocorrência (campo: *data_inversa*) dos registros dentro dos arquivos CSV, visando maior performance nas ocasionais consultas por data/hora e economia de bytes por transação, impactando assim no consumo de créditos da plataforma.

Para a criação da tabela particionada e carga dos dados, utilizei a ferramenta “*Cloud Shell*”, permitindo realizar estas operações através de comandos da plataforma conforme demonstrado abaixo:

- Criação da tabela particionada e especificação dos campos:

```
bq mk --table --time_partitioning_field=data_inversa --schema 'id:STRING, data_inversa:DATE,
dia_semana:STRING, horario:STRING, uf:STRING, br:STRING, km:STRING, municipio:STRING,
causa_acidente:STRING, tipo_acidente:STRING, classificacao_acidente:STRING, fase_dia:STRING,
sentido_via:STRING, condicao_meteorologica:STRING, tipo_pista:STRING, tracado_via:STRING,
uso_solo:STRING, pessoas:STRING, mortos:STRING, feridos_leves:STRING, feridos_graves:STRING,
ilecos:STRING, ignorados:STRING, feridos:STRING, veiculos:STRING, latitude:STRING,
longitude:STRING, regional:STRING, delegacia:STRING, uop:STRING' datatran_db.ocorrencias
```

- Carga dos arquivos CSV do Cloud Storage:

```
bq load --skip_leading_rows=1 --field_delimiter=";" --encoding="ISO-8859-1"
datatran_db.ocorrencias gs://rafaelferraz-storage/prf/datatran2020.csv,gs://rafaelferraz-
storage/prf/datatran2021.csv,gs://rafaelferraz-storage/prf/datatran2022.csv,gs://rafaelferraz-
storage/prf/datatran2023.csv
```



```

CLOUD SHELL
Terminal (daring-pilot-398315) x +
Abrir editor

git: [git version 2.30.2]
ssh: [OpenSSH_8.4p1 Debian-5+deb11u1, OpenSSL 1.1.1n 15 Mar 2022]

rafaelferraz df@cloudshell:~ (daring-pilot-398315)$ bq mk --table --time partitioning field=data_inversa --schema 'id:STRING, data_inversa:DATE, dia_semana:STRING, horari
oi:STRING, uf:STRING, br:STRING, km:STRING, municipio:STRING, causa_acidente:STRING, tipo_acidente:STRING, classificacao_acidente:STRING, fase_dia:STRING, sentido_via:STRI
NG, condicao_meteorologica:STRING, tipo_pista:STRING, tracado_via:STRING, uso_solo:STRING, pessoas:STRING, mortos:STRING, feridos_leves:STRING, feridos_graves:STRING, ile
sos:STRING, ignorados:STRING, feridos:STRING, veiculos:STRING, latitude:STRING, longitude:STRING, regional:STRING, delegacia:STRING, uop:STRING' datatran_db.ocorrencias
Table 'daring-pilot-398315:datatran_db.ocorrencias' successfully created.
rafaelferraz df@cloudshell:~ (daring-pilot-398315)$ bq load --skip_leading_rows=1 --field_delimiter=";" --encoding="ISO-8859-1" datatran_db.ocorrencias gs://rafaelferraz-
storage/prf/datatran2020.csv,gs://rafaelferraz-storage/prf/datatran2021.csv,gs://rafaelferraz-storage/prf/datatran2022.csv,gs://rafaelferraz-storage/prf/datatran2023.csv
Waiting on bqjob r21a027ded0a1d672 000001ba3f46590 1 ... (14s) Current status: DONE
rafaelferraz df@cloudshell:~ (daring-pilot-398315)$

```

Figura 2- Evidência da execução dos comandos.

Ao fim da carga, foram contabilizados um total de 236.697 linhas referentes aos arquivos com os dados históricos de 2020 até 2023:

Aqui estão os recursos do espaço de trabalho.

MOstrar APENAS COM ESTRELA

- daring-pilot-398315
 - Conexões externas
 - datatran_db
 - ocorrencias**
 - temp_dataset_beam_bq_jo...
 - temp_dataset_beam_bq_jo...
 - temp_dataset_beam_bq_jo...

ocorrencias

CONSULTA COMPARTILHAR

Esta é uma tabela **particionada**. [Learn more](#)

ESQUEMA DETALHES PREVIEW LINHAGEM PERF

Informações do armazenamento

Número de linhas	236.697
Número de partições	0
Total de bytes lógicos	64,84 MB
Bytes lógicos ativos	64,84 MB
Bytes lógicos de longo prazo	0 B
Total de bytes físicos	0 B
Bytes físicos ativos	0 B
Bytes físicos de longo prazo	0 B
Bytes físicos de viagem no tempo	0 B

Figura 3- Detalhamento da tabela de "ocorrencias".

ocorrencias

CONSULTA COMPARTILHAR COPIAR SNAPSHOT EXCLUIR ATUALIZAR

Esta é uma tabela particionada. [Learn more](#)

DISMISS

ESQUEMA	DETALHES	PREVIEW	LINHAGEM	PERFIL DE DADOS	QUALIDADE DOS DADOS
Linha	causa_acidente	tipo_acidente	classificacao_acidente	fase_dia	sentido_via
1	Curva acentuada	Saída de leito carroçável	Com Vítimas Feridas	Pleno dia	Crescente
2	Mal súbito do condutor	Saída de leito carroçável	Com Vítimas Feridas	Pleno dia	Crescente
3	Condutor Dormindo	Saída de leito carroçável	Com Vítimas Feridas	Plena Noite	Crescente
4	Avarias e/ou desgaste excessi...	Tombamento	Com Vítimas Feridas	Amanhecer	Decrescente
5	Entrada inopinada do pedestre	Atropelamento de Pedestre	Com Vítimas Feridas	Pleno dia	Decrescente
6	Iluminação deficiente	Saída de leito carroçável	Com Vítimas Feridas	Plena Noite	Decrescente
7	Ultrapassagem Indevida	Saída de leito carroçável	Com Vítimas Feridas	Pleno dia	Crescente
8	Condutor Dormindo	Tombamento	Com Vítimas Feridas	Plena Noite	Crescente
9	Entrada inopinada do pedestre	Atropelamento de Pedestre	Com Vítimas Fatais	Plena Noite	Crescente
10	Mal súbito do condutor	Colisão com objeto	Com Vítimas Feridas	Plena Noite	Crescente

Resultados por página: 50 1 – 50 de 236697

Figura 4- Detalhamento da tabela de “ocorrencias”.

6.2. Modelagem no Cloud SQL

Nesta etapa optei por utilizar o serviço de banco de dados relacional *Cloud SQL* da plataforma, criando uma instância denominada “dw-datatran” do MySQL para implantação do modelo de banco de dados em estrela.

SQL Instâncias CRIAR INSTÂNCIA MIGRAR DADOS MOSTRAR PAINEL DE INFORMAÇÕES

Filtro Insira o nome ou o valor da propriedade

ID da instância	Edição do Cloud SQL	Tipo	Endereço IP público	Endereço IP particular	Nome da conexão da instância	Alta disponibilidade	Local	Ações
dw-datatran	Enterprise	MySQL 8.0	34.95.231.77		daring-pilot-398315.sout...	ATIVAR	southern	

Figura 5- Instância ativa do MySQL no Cloud SQL.

Visão geral EDITAR IMPORTAR EXPORTAR REINICIAR INTERROMPER EXCLUIR CLONAR

Conectar-se a esta instância

Endereço IP público

34.95.231.77

Nome da conexão

daring-pilot-398315:southamerica-east1:dw-datatran

Precisa de ajuda com a conexão?

Leia a documentação para saber mais sobre as diversas maneiras de se conectar à instância. [Saiba mais](#)

Para se conectar usando o gcloud, [ABRIR O CLOUD SHELL](#)

Para saber mais sobre como se conectar a uma VM do Compute Engine, [INICIAR TUTORIAL](#)

Ações sugeridas

→ Criar backup

Configuração

vCPUs	Memória	Armazenamento SSD
1	3,75 GB	10 GB

- Edição Enterprise
- A versão do banco de dados é MySQL 8.0.31
- O aumento automático do armazenamento está ativado
- Os backups automáticos estão desativados
- A recuperação pontual está desativada.
- A proteção contra exclusão de instâncias está desativada
- Localizada em southamerica-east1-c
- Não é altamente disponível (por zona)
- Nenhuma flag do banco de dados foi configurada
- Nenhum identificador foi configurado

Figura 6 - Configuração da instância.

Com base na observação dos atributos dos arquivos coletados e distinção dos campos para definição das dimensões e fatos, foi elaborado o desenho do modelo de banco de dados denominado “datatran” conforme modelo abaixo:

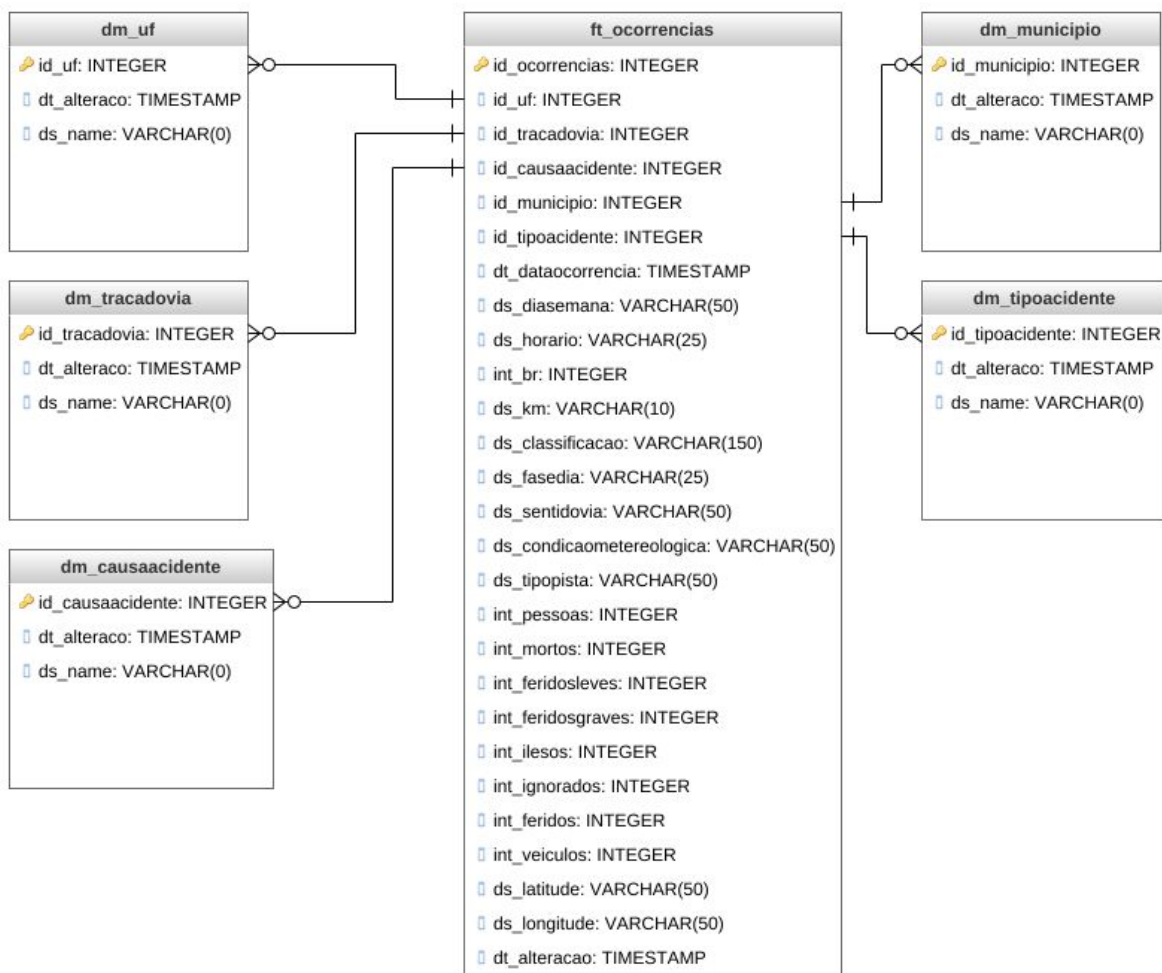


Figura 7- Diagrama do banco "datatran" desenhado com a ferramenta GenMyModel.

Para uma maior clareza na identificação dos campos ao realizar consultas, foram definidos os seguintes prefixos nos nomes dos atributos que remetem ao tipo do dado armazenado:

id_ = Chaves primárias e estrangeiras

ds_ = Texto

dt_ = Data e/ou hora

int_ = Inteiro

dm_ = Tabela de dimensão

ft_ = Tabela de fato

Além disso, todas as tabelas recebem um campo de data e hora denominado “dt_alteracao”, constando a informação da última operação de escrita realizada pela ferramenta de ETL naquele registro.

Com base no modelo de dados desenvolvido na figura 7, foi elaborado a construção de um script SQL para criação das tabelas, seus respectivos atributos, tipos dos dados e chaves. A ferramenta DBeaver (<https://dbeaver.io>) foi escolhida para conectar-se externamente à instância de banco de dados da nuvem e processar os scripts SQL utilizados neste trabalho.

Observação: O script (*script_dw.sql*) desenvolvido para esta etapa está disponível junto com o material deste MVP no GitHub.

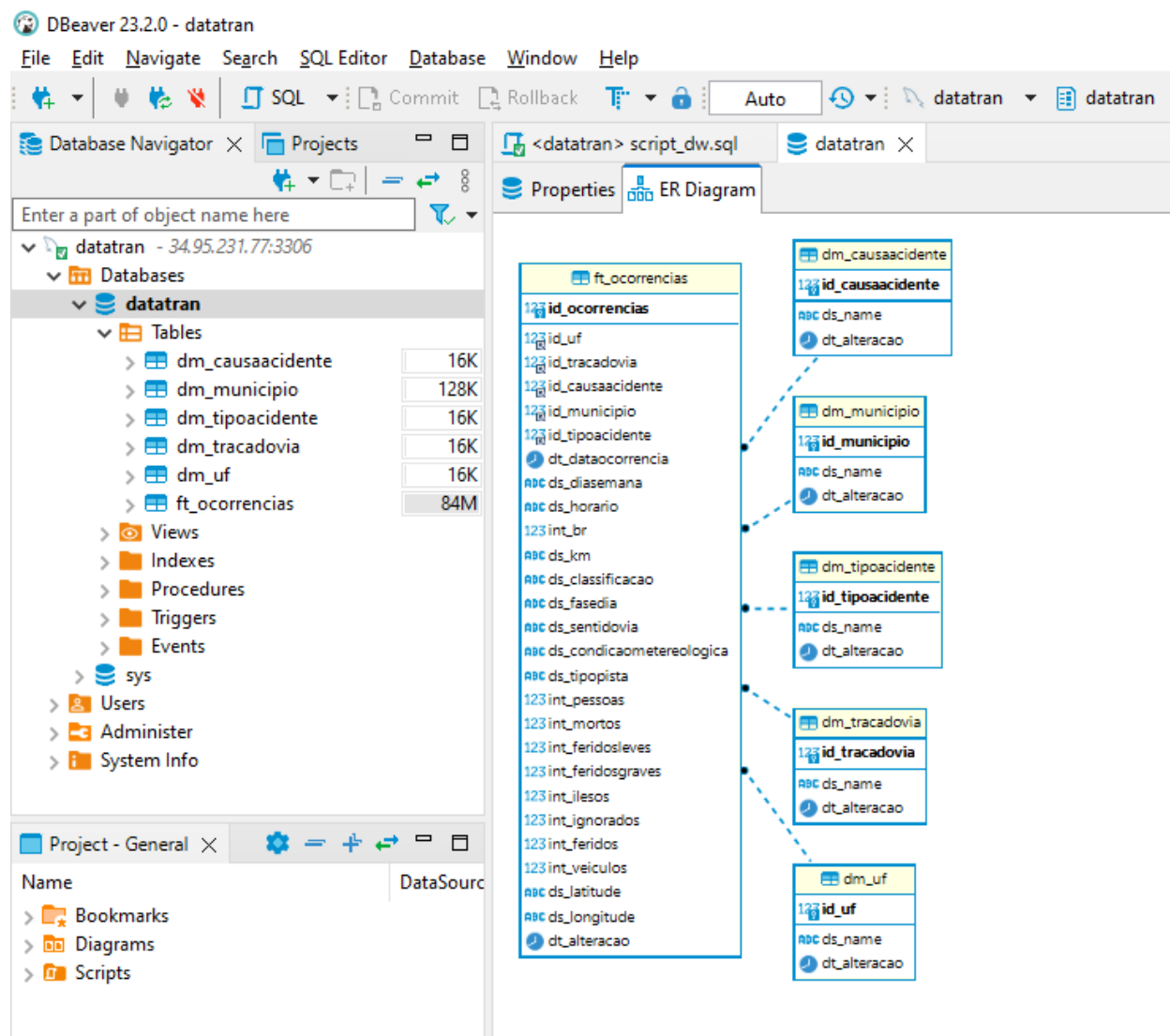


Figura 8- Modelo de dados carregado no DBeaver.

7. Catálogo de Dados

Para uma maior elucidação do modelo de dados, abaixo seguem os detalhes de cada tabela do banco de dados “datatran”.

Observação: Segundo a documentação oficial, para os campos de “id” (inteiro) o valor máximo atribuído é de 2147483647.

Fonte: <https://dev.mysql.com/doc/refman/8.0/en/integer-types.html>

Tabela: dm_causaacidente

Visão Geral: Relação das causas de acidentes catalogadas pelas ocorrências.

Campo	Tipo	Nulo?	Chave?	Domínio	Descrição
id_causaacidente	Inteiro	Não	Sim	0-2147483647	Chave primária
ds_name	Texto	Não	Não	255 caracteres	Causa do acidente
dt_alteracao	Data e Hora	Não	Não	AAAA-MM-DD HH:MI:SS	Data da última alteração

Tabela 1 - dm_causaacidente

Tabela: dm_municipio

Visão Geral: Relação dos municípios presentes nas ocorrências de acidentes.

Campo	Tipo	Nulo?	Chave?	Domínio	Descrição
id_municipio	Inteiro	Não	Sim	0-2147483647	Chave primária
ds_name	Texto	Não	Não	255 caracteres	Nome do município
dt_alteracao	Data e Hora	Não	Não	AAAA-MM-DD HH:MI:SS	Data da última alteração

Tabela 2 - dm_municipio

Tabela: dm_tipoacidente

Visão Geral: Relação dos tipos de acidentes presentes nas ocorrências de acidentes.

Campo	Tipo	Nulo?	Chave?	Domínio	Descrição
id_tipoacidente	Inteiro	Não	Sim	0-2147483647	Chave primária
ds_name	Texto	Não	Não	255 caracteres	Tipo do acidente
dt_alteracao	Data e Hora	Não	Não	AAAA-MM-DD HH:MI:SS	Data da última alteração

Tabela 3 - dm_tipoacidente

Tabela: dm_tracadovia

Visão Geral: Relação dos traçados de vias aonde ocorreram os acidentes.

Campo	Tipo	Nulo?	Chave?	Domínio	Descrição
id_tracadovia	Inteiro	Não	Sim	0-2147483647	Chave primária
ds_name	Texto	Não	Não	255 caracteres	Traçado da via
dt_alteracao	Data e Hora	Não	Não	AAAA-MM-DD HH:MI:SS	Data da última alteração

Tabela 4 - dm_tracadovia

Tabela: dm_uf

Visão Geral: Relação das Unidades Federativas (UFs) aonde ocorreram os acidentes.

Campo	Tipo	Nulo?	Chave?	Domínio	Descrição
id_uf	Inteiro	Não	Sim	0-2147483647	Chave primária
ds_name	Texto	Não	Não	255 caracteres	Nome da Unidade Federativa
dt_alteracao	Data e Hora	Não	Não	AAAA-MM-DD HH:MI:SS	Data da última alteração

Tabela 5 - dm_uf

Tabela: ft_ocorrencias

Visão Geral: Registros detalhados de ocorrências de acidentes registrados pela Polícia Rodoviária Federal referente ao período de 2020 até 2023.

Campo	Tipo	Nulo?	Chave?	Domínio	Descrição
id_ocorrencias	Inteiro	Não	Sim	0-2147483647	Chave primária
id_uf	Inteiro	Não	Sim	0-2147483647	Chave estrangeira para a tabela de UF
id_tracadovia	Inteiro	Não	Sim	0-2147483647	Chave estrangeira para a tabela de Traçado Via
id_causaacidente	Inteiro	Não	Sim	0-2147483647	Chave estrangeira para a tabela de Causa Acidente
id_municipio	Inteiro	Não	Sim	0-2147483647	Chave estrangeira para a tabela de Município
id_tipoacidente	Inteiro	Não	Sim	0-2147483647	Chave estrangeira para a tabela de Tipo Acidente
dt_dataocorrencia	Data	Não	Não	AAAA-MM-DD	Data da ocorrência do acidente
ds_diasemana	Texto	Não	Não	Segunda-feira, terça-feira, quarta-feira, quinta-feira, sexta-feira, sábado e domingo	Dia da semana em que ocorreu o acidente
ds_horario	Texto	Sim	Não	1970-01-01 Hh:mm:ss	Horário em que ocorreu o acidente
int_br	Inteiro	Sim	Não	0-999	Número da BR onde ocorreu o acidente
ds_km	Texto	Sim	Não	0-999,9, NA	Quilômetro da estrada onde ocorreu o acidente
ds_classificacao	Texto	Não	Não	Com Vítimas Feridas, Com Vítimas Fatais, Sem Vítimas	Classificação do acidente
ds_fasedia	Texto	Não	Não	Pleno dia, Plena Noite, Amanhecer, Anoitecer	Fase do dia em que ocorreu o acidente

Campo	Tipo	Nulo?	Chave?	Domínio	Descrição
ds_sentidovia	Texto	Não	Não	Crescente, Decrescente, Não Informado	Sentido da via em que ocorreu o acidente
ds_condicaometereologica	Texto	Não	Não	Céu Claro, Nublado, Chuva, Garoa/Chuveiro, Sol, Ignorado, Vento, Nevoeiro/Neblina, Granizo, Neve	Condição meteorológica no momento em que ocorreu o acidente
ds_tipopista	Texto	Não	Não	Dupla, Simples, Múltipla	Tipo da pista em que ocorreu o acidente
int_pessoas	Inteiro	Não	Não	0-99999999	Quantidade de pessoas envolvidos no acidente
int_mortos	Inteiro	Não	Não	0-99999999	Quantidade de pessoas que vieram a óbito
int_feridosleves	Inteiro	Não	Não	0-99999999	Quantidade de pessoas que se feriram levemente no acidente
int_feridosgraves	Inteiro	Não	Não	0-99999999	Quantidade de pessoas que se feriram gravemente no acidente
int_ilesos	Inteiro	Não	Não	0-99999999	Quantidade de pessoas que ficaram ilesas no acidente
int_ignorados	Inteiro	Não	Não	0-99999999	Quantidade de pessoas ignoradas no acidente
int_feridos	Inteiro	Não	Não	0-99999999	Quantidade de pessoas feridas no acidente
int_veiculos	Inteiro	Não	Não	0-99999999	Quantidade de veículos envolvidos no acidente
ds_latitude	Texto	Não	Não	-90,000000 até 90,000000	Coordenada geográfica de latitude onde ocorreu o acidente

Campo	Tipo	Nulo?	Chave?	Domínio	Descrição
ds_longitude	Texto	Não	Não	-180,000000 até 180,000000	Coordenada geográfica de longitude onde ocorreu o acidente
dt_alteracao	Data e Hora	Não	Não	AAAA-MM-DD HH:MI:SS	Data da última alteração do registro pelo

Tabela 6 - ft_ocorrencias

8. Carga

Para o procedimento de carga do modelo de dados em estrela, optei por explorar os recursos da ferramenta de ETL “Cloud Dataprep by Trifacta”, um serviço inteligente de dados em nuvem para exploração visual, limpeza e preparação de dados para análise, integrado à plataforma do Google Cloud.

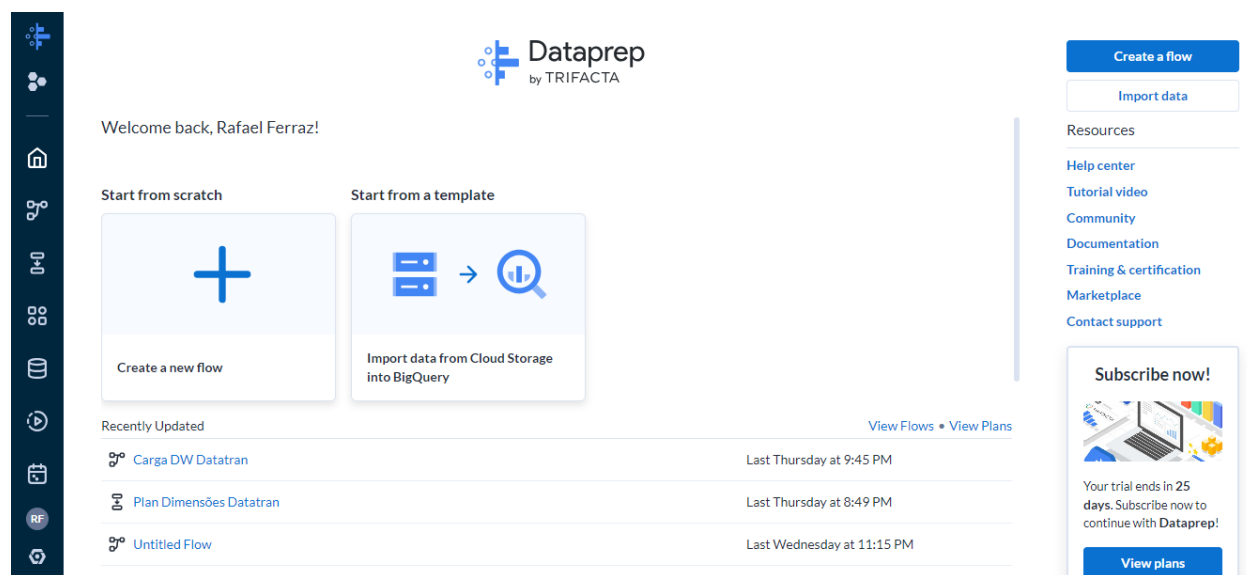


Figura 9- Interface da página inicial do Dataprep.

8.1. Etapas da Carga

Os dados brutos (oriundos da tabela de ocorrências do BigQuery) foram tratados e destinados às suas respectivas tabelas do modelo relacional, respeitando os tipos de atributo e integridade das chaves e relacionamentos entre as tabelas. Para isso, a carga foi realizada em duas etapas e na seguinte sequência: Carga das tabelas de **Dimensão** e Carga da tabela de **Fato**, conforme detalhado abaixo.

8.1.1. Carga das Dimensões

Foi criado um fluxo denominado “Carga DW Datatran”, que contempla a carga de todo o modelo estrela na instância do MySQL. Em destaque abaixo (linhas em azul), está a carga das tabelas de **dimensão** do modelo, aonde os dados de cada uma das 5 dimensões são extraídos do BigQuery (Dataset), agrupados, tipados, datados, sequenciados (Recipe) e por fim, carregados em suas respectivas tabelas de domínio (Output).

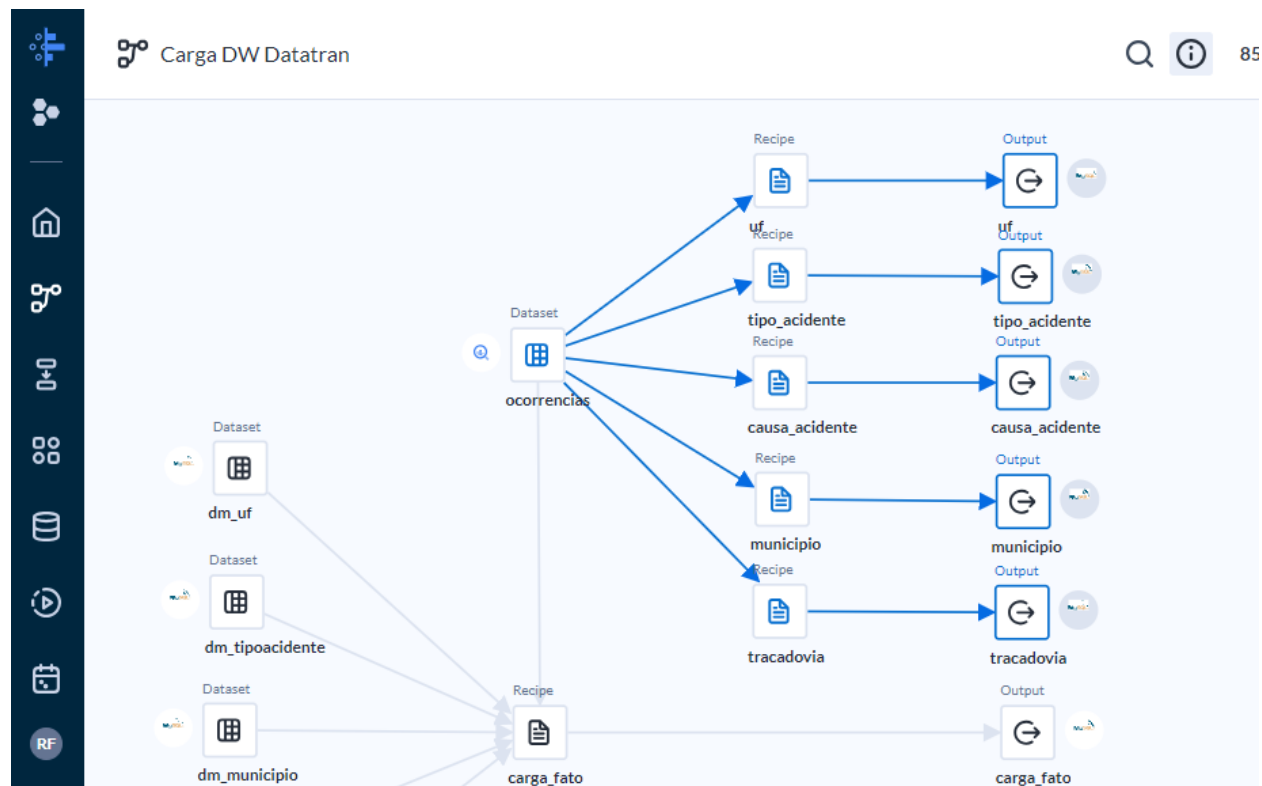


Figura 10- Em destaque, as etapas da carga das dimensões

Para fins de demonstração, serão detalhadas as configurações do fluxo de carga da tabela de UFN. Para as demais dimensões (Tipo Acidente, Causa Acidente, Município e Traçado Via) os procedimentos são os mesmos, mudando apenas o atributo utilizado.

Na etapa de leitura dos dados (ícone de “Dataset”), foi feito o apontamento para a tabela Flat de ocorrências no BigQuery conforme detalhado abaixo:

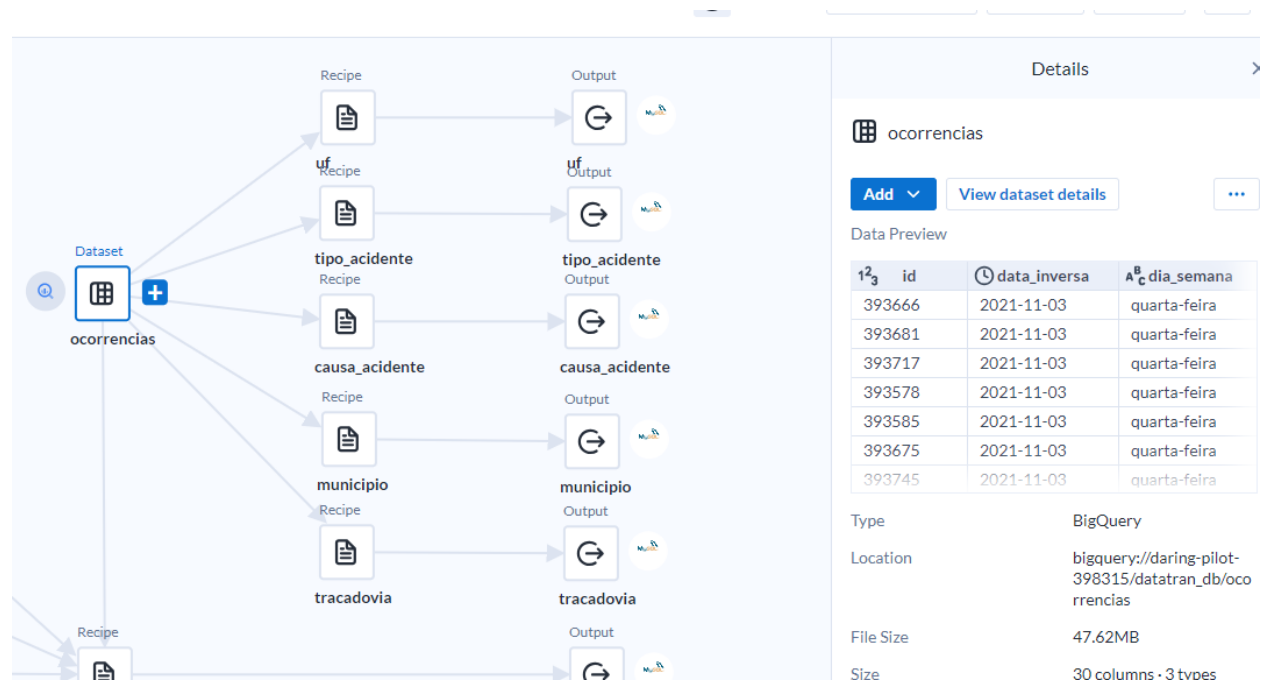
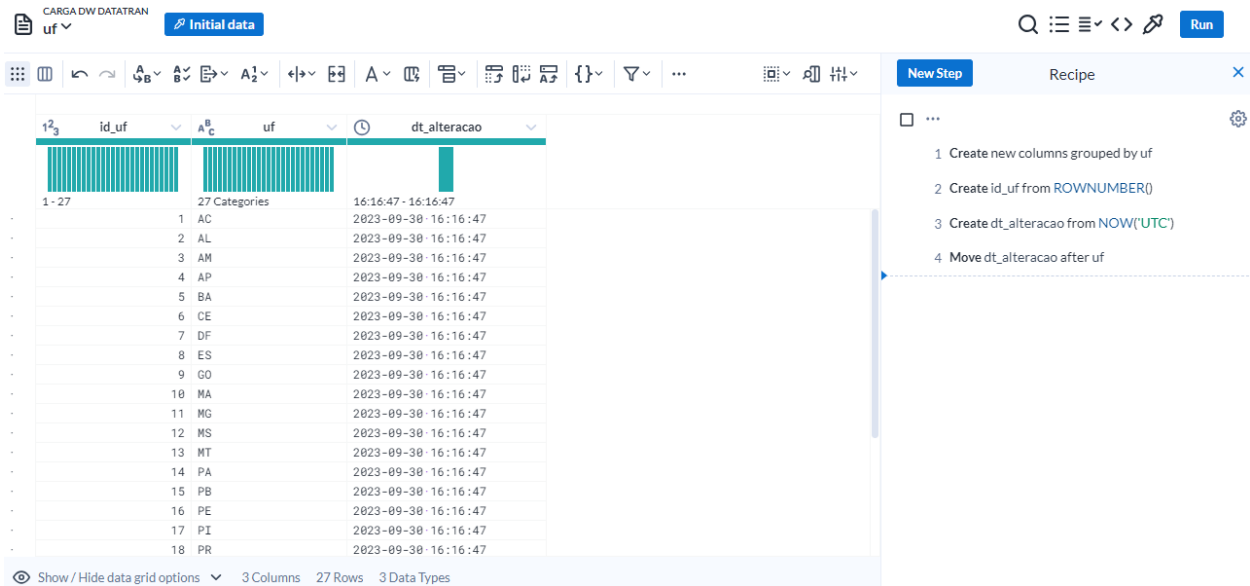


Figura 11- Leitura dos dados

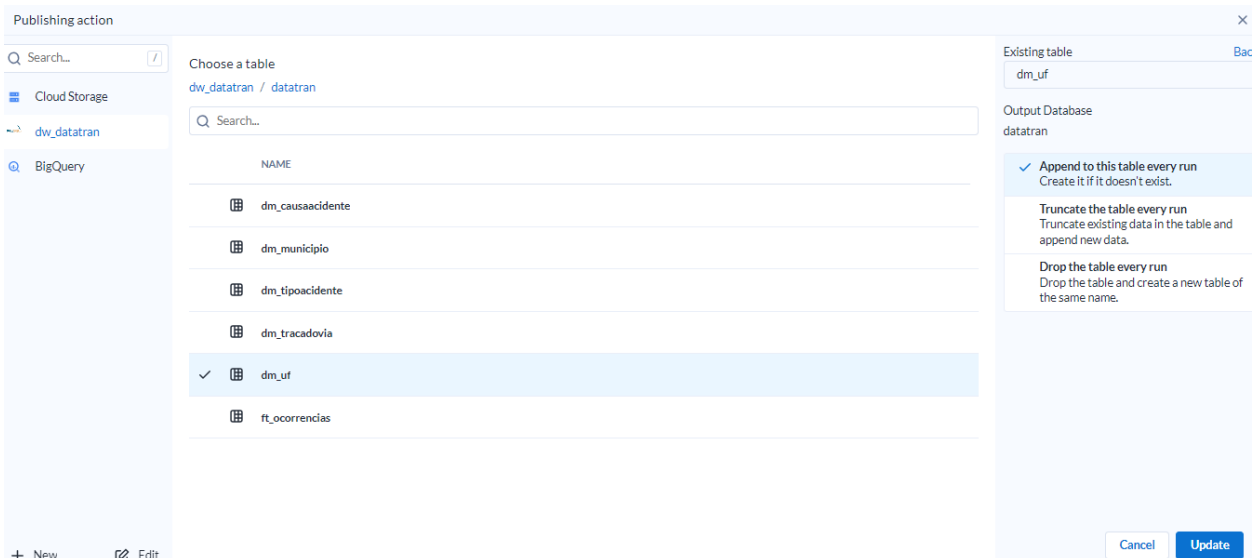
Em seguida, é iniciada a etapa de transformação (Recipe), onde o campo UF é agrupado para distinguir as unidades. Adicionalmente são incluídas as colunas de id (sequencial automático) e data de alteração (momento em que a carga é processada). Na imagem abaixo, estas ações ficam evidentes no painel do lado direito, enquanto no lado esquerdo é mostrada uma prévia dos dados após a transformação.



id_uf	uf	dt_alteracao
1	AC	2023-09-30 16:16:47
2	AL	2023-09-30 16:16:47
3	AM	2023-09-30 16:16:47
4	AP	2023-09-30 16:16:47
5	BA	2023-09-30 16:16:47
6	CE	2023-09-30 16:16:47
7	DF	2023-09-30 16:16:47
8	ES	2023-09-30 16:16:47
9	GO	2023-09-30 16:16:47
10	MA	2023-09-30 16:16:47
11	MG	2023-09-30 16:16:47
12	MS	2023-09-30 16:16:47
13	MT	2023-09-30 16:16:47
14	PA	2023-09-30 16:16:47
15	PB	2023-09-30 16:16:47
16	PE	2023-09-30 16:16:47
17	PI	2023-09-30 16:16:47
18	PR	2023-09-30 16:16:47

Figura 12 - Interface de transformação dos dados da tabela dm_uf

Por fim, é realizada a carga (Output) na tabela de UF no modelo relacional do MySQL. Destaque para o painel no lado esquerdo onde são exibidas as conexões com outras ferramentas elegíveis do Google Cloud, o painel central com as tabelas da conexão selecionada e o painel esquerdo para a escolha de como será feita a carga: Incrementar e Adicionar, Limpar antes de carregar e apagar a tabela/criar novamente.



Publishing action

Search...

Cloud Storage

dw_datatran

BigQuery

Choose a table

dw_datatran / datatran

Search...

NAME
dm_causaacadente
dm_municipio
dm_tipoacadente
dm_tracadovia
✓ dm_uf
ft_ocorrencias

Existing table

dm_uf

Output Database

datatran

✓ Append to this table every run
Create it if it doesn't exist.

Truncate the table every run
Truncate existing data in the table and append new data.

Drop the table every run
Drop the table and create a new table of the same name.

Cancel Update

Figura 13 - Interface de conexão de saída dos dados

Ainda na etapa que antecede a escolha da conexão, o Dataprep permite escolher o ambiente aonde será processada a carga dos dados, sendo no “Trifacta Photon” da própria ferramenta, recomendado para cargas leves (de até 1GB de dados processados) e no “Dataflow” para cargas que exigem mais processamento. Para todas as cargas desse fluxo, o “Trifacta Photon” foi satisfatório.

Run Job

Running Environment

- ☒ **Trifacta Photon**
Run job on Trifacta Photon (best for small and medium-sized jobs, up to approximately 1 GB of data)
- ☐ **Dataflow**
Run job on Dataflow

Options

- ☒ **Profile results and assess data quality rules**
Generate a statistical profile of the published data and evaluate data quality rules
- ☒ **Validate schema**
Analyze input schema and compare with the flow dataset schema.
Currently applicable to relational datasets and limited file types. [Learn more.](#)
- ☐ **Fail job if dataset schemas change**
Fail the job if the input data schema differs from the schema of the flow dataset.
- ☐ **Ignore recipe errors**
Allow jobs to be run even if there are errors present in recipe steps

Publishing Actions

Actions	Location	Settings
Append-Cloud_mysql	Connection: dw_datatran; Database: datatran; Table: dm_uf	Create table if it does not exist; append

Figura 14- Tela inicial de configuração do Output.

Ainda na tela de configuração do Output, também é possível escolher um script SQL que será executado antes ou após a carga:

Publishing Actions

Actions	Location	Settings
Append-Cloud_mysql	Connection: dw_datatran; Database: datatran; Table: dm_uf	Create table if it does not exist; append

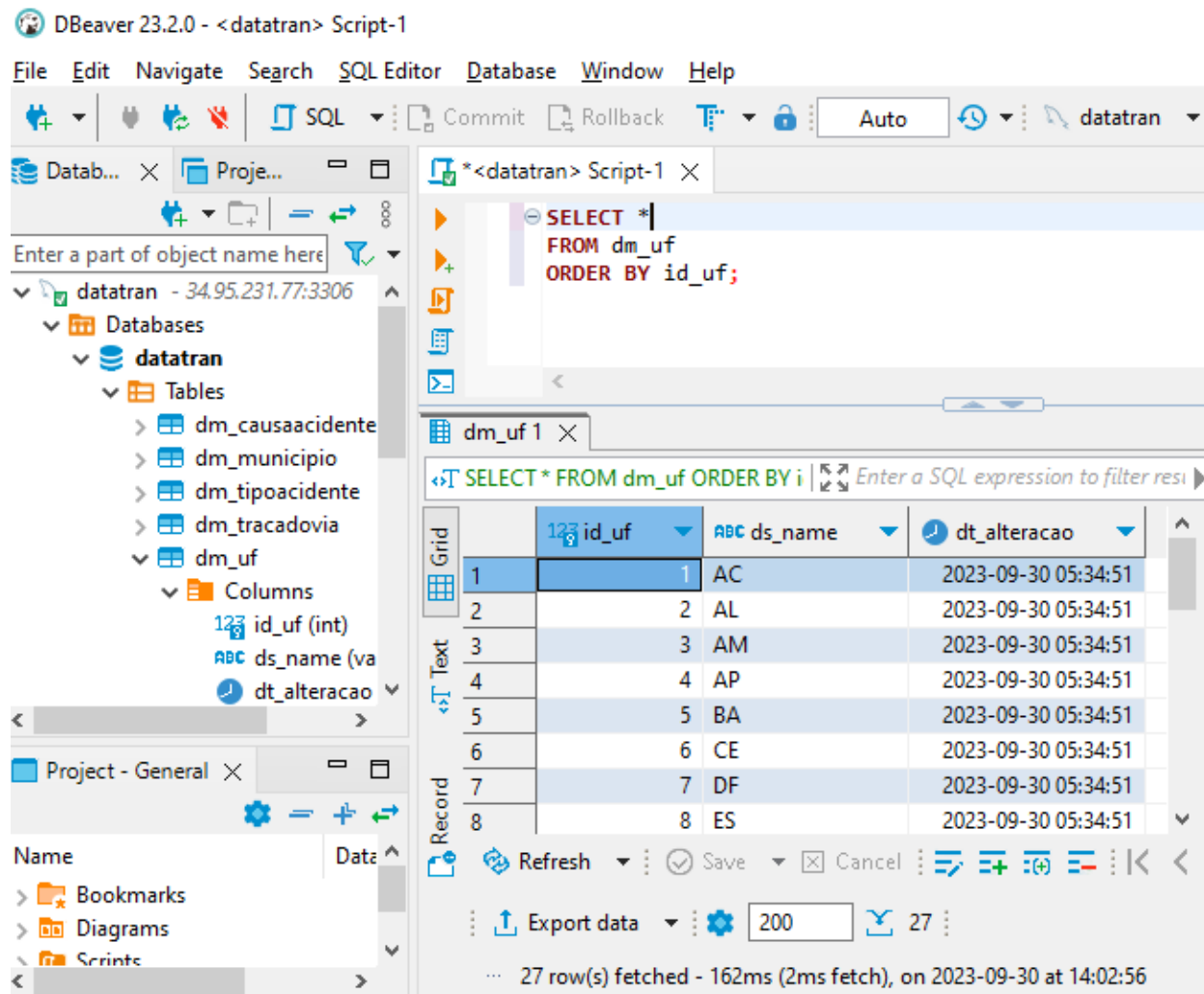
SQL Scripts

Connection	SQL statement	Settings
dw_datatran	-- limpa a tabela fato DELETE FROM datatran.ft_ocorrencias; -- limpa as tabelas de dimensao DELETE FROM datatran.dm_uf; DELETE FROM datatran.dm_municipio; DELETE FROM datatran.dm_tipoacidente; DELETE FROM datatran.dm_tracadovia; DELETE FROM datatran.dm_causa...	Run before data ingest

Figura 15- Opção para incluir um script SQL antes ou após a carga.

Observação: Para a etapa de carga da UF em específico, por ser a primeira a correr antes de todas as outras, eu optei por utilizar um script (e apenas nesta etapa) que limpa todas as tabelas do modelo estrela antes de iniciar as cargas como “solução de contorno” devido a problemas de integridade de chave enfrentados durante as cargas. Optei por esta ação devido ao curto tempo para implantação da solução e mais a frente abordarei esse assunto no capítulo Melhoria Futura.

Ao fim da carga, os dados já podem ser consultados através da instância de MySQL. No meu caso utilizei a ferramenta DBeaver, conectando-se externamente à instância através de minha máquina local:



The screenshot shows the DBeaver 23.2.0 interface. The left sidebar displays the database structure for 'datatran' at IP 34.95.231.77:3306, including tables like 'dm_causaacidente', 'dm_municipio', 'dm_tipoacidente', 'dm_tracadovia', and 'dm_uf'. The 'dm_uf' table is selected, showing columns 'id_uf (int)', 'ds_name (varchar)', and 'dt_alteracao (timestamp)'. The main window shows a SQL query: `SELECT * FROM dm_uf ORDER BY id_uf;`. Below the query, the results are displayed in a table grid with 8 rows and 3 columns: 'id_uf', 'ds_name', and 'dt_alteracao'. The first row is highlighted in blue.

	id_uf	ds_name	dt_alteracao
1	1	AC	2023-09-30 05:34:51
2	2	AL	2023-09-30 05:34:51
3	3	AM	2023-09-30 05:34:51
4	4	AP	2023-09-30 05:34:51
5	5	BA	2023-09-30 05:34:51
6	6	CE	2023-09-30 05:34:51
7	7	DF	2023-09-30 05:34:51
8	8	ES	2023-09-30 05:34:51

At the bottom of the results grid, it indicates '27 row(s) fetched - 162ms (2ms fetch), on 2023-09-30 at 14:02:56'.

Figura 16- Consulta na tabela dm_uf pelo DBeaver.

8.1.2. Carga da Fato

Para a carga da tabela Fato (linhas destacadas em azul), é necessário fazer a leitura das tabelas de dimensão carregadas na etapa anterior e coletar as chaves de cada uma das dimensões, atribuindo-as ao registro de ocorrências como “chave estrangeira”:

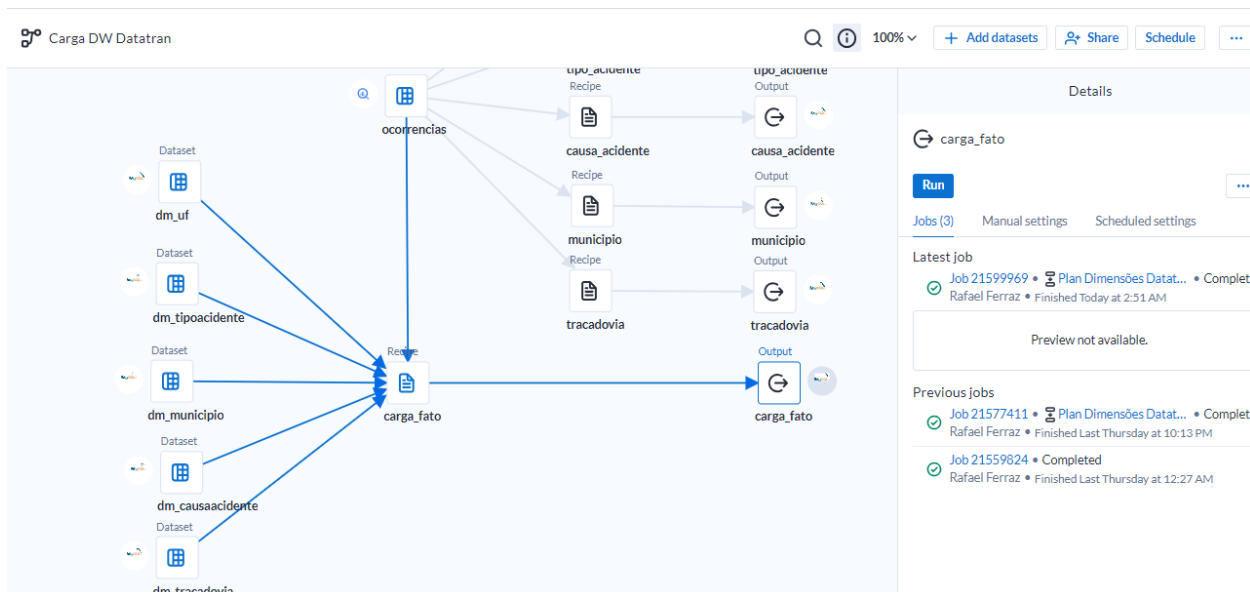


Figura 17- Fluxo de carga da tabela Fato.

Destaque para a etapa de transformação (Recipe) de “carga_fato”, aonde as entradas (Datasets) são cada uma das tabelas de dimensão. É crucial que esta etapa seja executada apenas após a conclusão bem-sucedida das tabelas de dimensão, onde darei mais detalhes no próximo capítulo “Plano de Execução e Agendamento”.

Na imagem abaixo, note que são executadas diversas transformações para consolidar a tabela fato (um total de 31 transformações). Essas transformações consistem na identificação do atributo no qual se deve buscar a chave nas tabelas de dimensão (Lookup), substituição desse atributo pela chave encontrada, ordenação dos atributos para se adequarem ao formato da tabela fato, além da inclusão das colunas de sequencial e data de alteração.

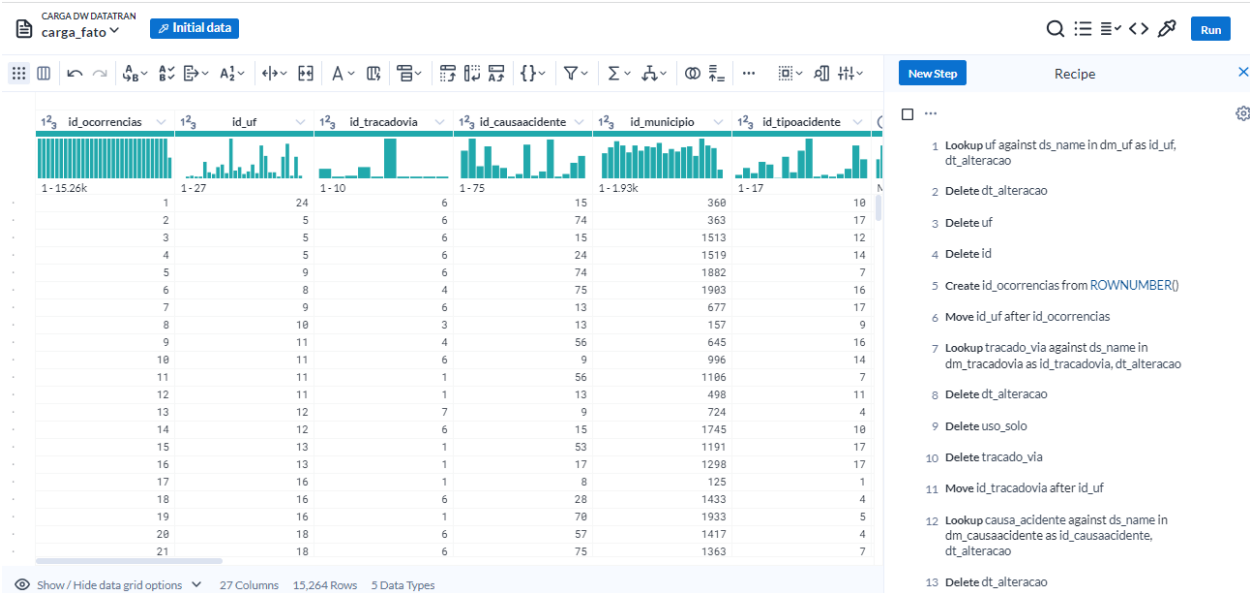


Figura 18- Configuração do recipe “carga_fato”, com as transformações aplicadas.

Observação: Uma característica identificada na ferramenta Dataprep é que ela não permite o mapeamento dos campos no momento de gravar a tabela. Logo espera-se que na etapa de transformação a saída seja fidedigna em relação à sequência dos campos e seus respectivos tipos de dados com a tabela de Output. Isso vale para a gravação das tabelas de dimensão e de fato.

Outra característica observada na ferramenta é que ao escolher um atributo para fazer a busca de sua chave de referência através da opção “Lookup” (imagem à esquerda), os datasets de entrada são atribuídos automaticamente à transformação (imagem à direita):

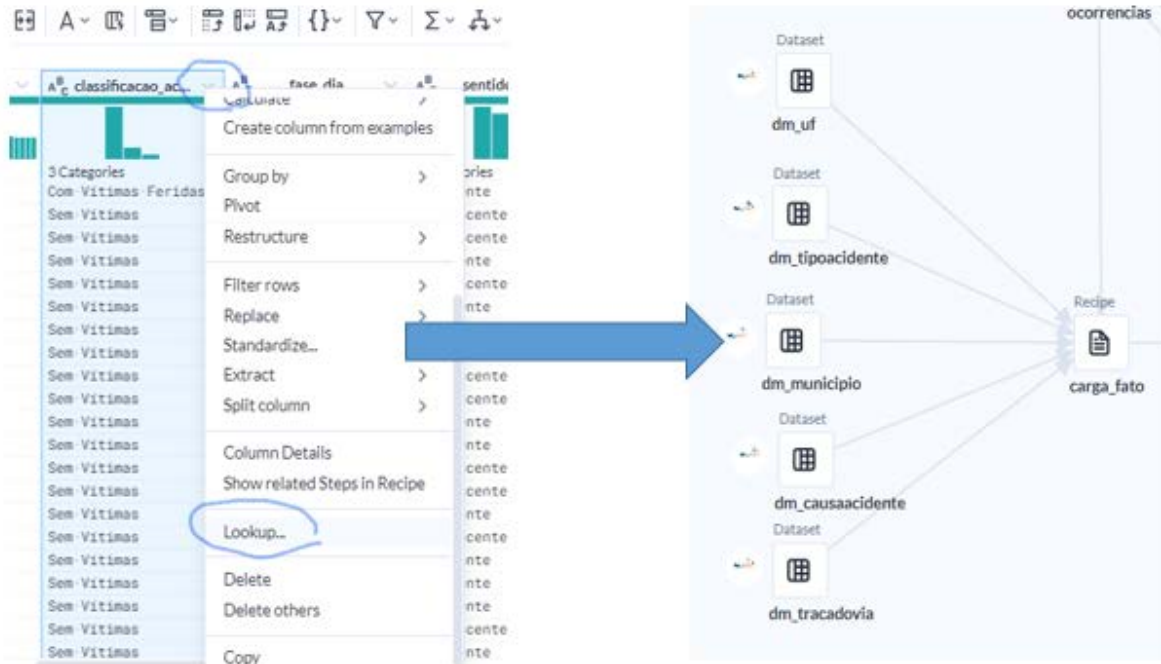


Figura 19 - Comportamento da opção "Lookup" no Dataprep

Ao fim da execução bem-sucedida, foi constatado a carga dos dados na tabela "ft_ocorrencias" no modelo de dados estrela, com os valores das chaves devidamente referenciados com as tabelas de dimensão:

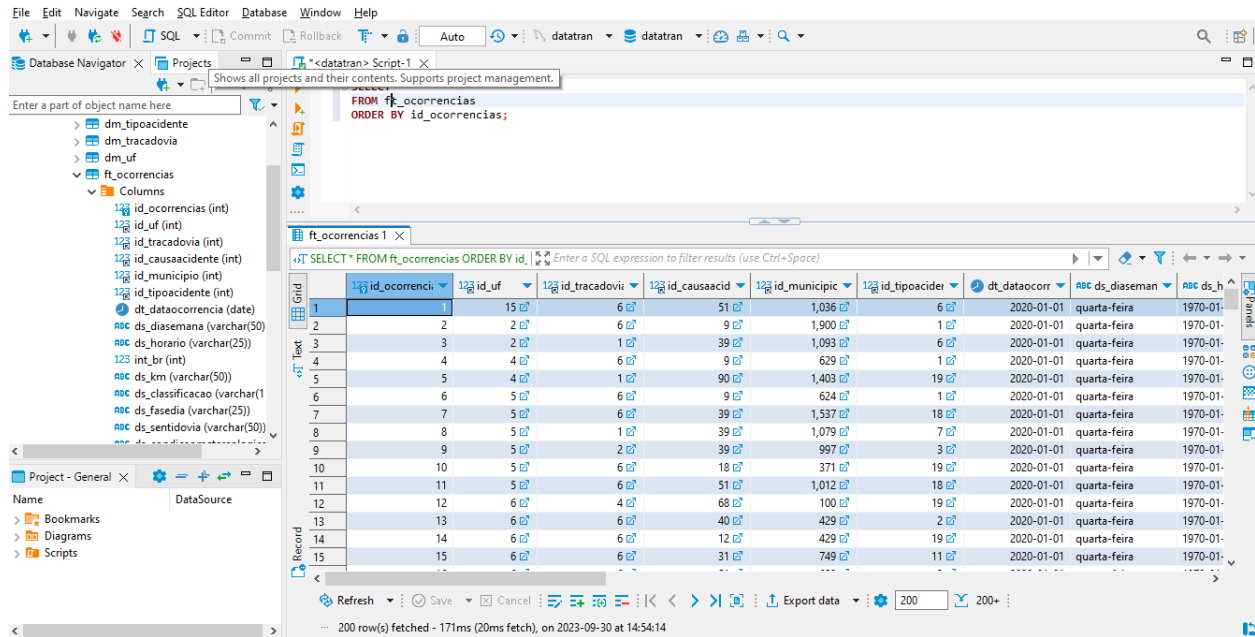


Figura 20- Consulta à tabela de `ft_ocorrencias` no DBeaver.

9. Plano de Execução e Agendamento

Para que seja respeitada a sequência de execução das cargas das tabelas de Dimensão e Fato, é necessário que haja um orquestrador para conduzir as cargas e, em caso de falha, abortar a sequência de carga para as próximas etapas. Essa etapa é crucial para que a integridade de dados entre as dimensões e a fato sejam satisfatórias. Para isso foi utilizado o módulo “Plans” do próprio Dataprep:

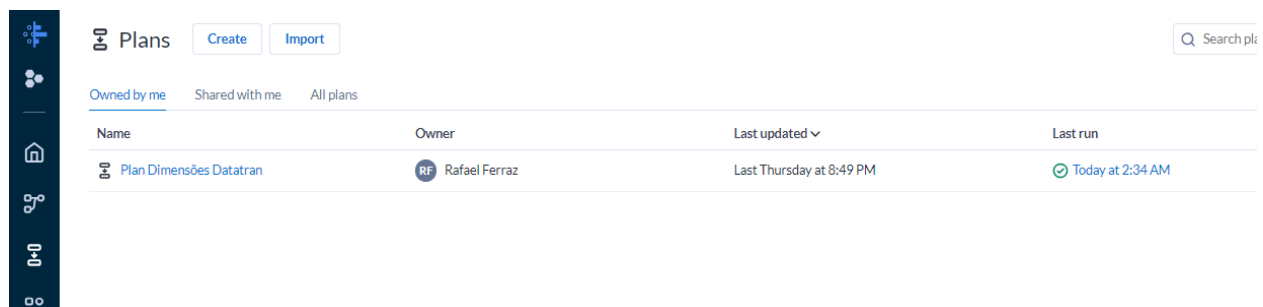


Figura 21- Planos de execução no módulo Plans.

No plano de execução, foi realizada a configuração para que a carga da tabela fato seja executada apenas após a execução bem-sucedida das dimensões:

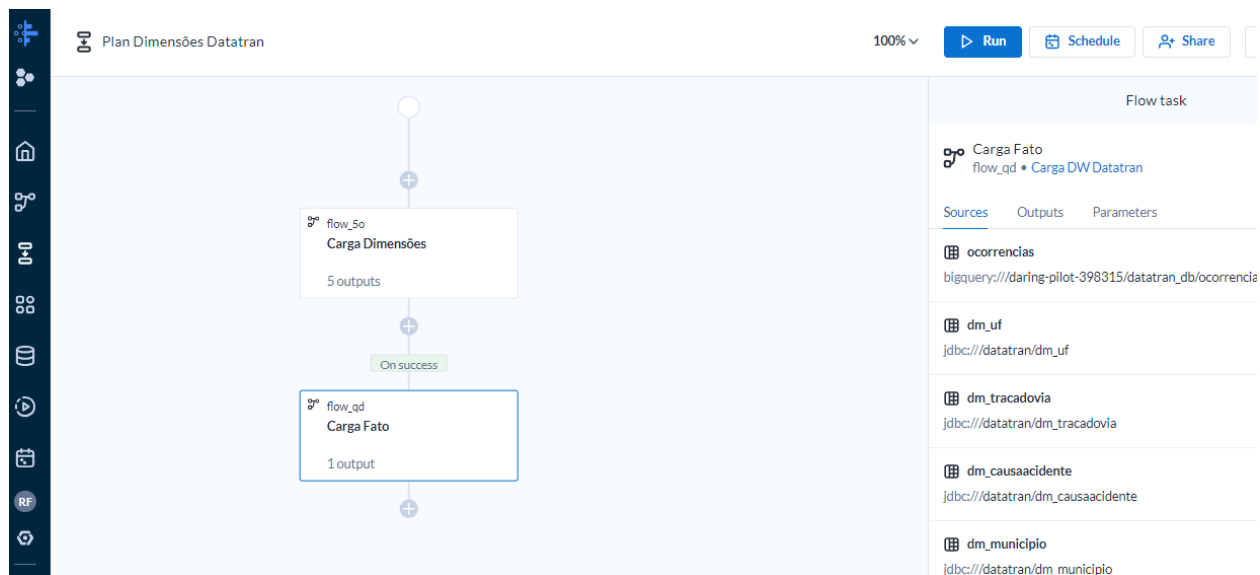


Figura 22- Sequência de execução. (De cima para baixo)

Através do módulo “Schedules” da ferramenta é possível configurar um agendamento para que o fluxo seja executado automaticamente, com o intuito de manter o esquema de dados sempre atualizado. Neste projeto eu defini a execução do plano para ocorrer semanalmente aos domingos, sempre à meia-noite:

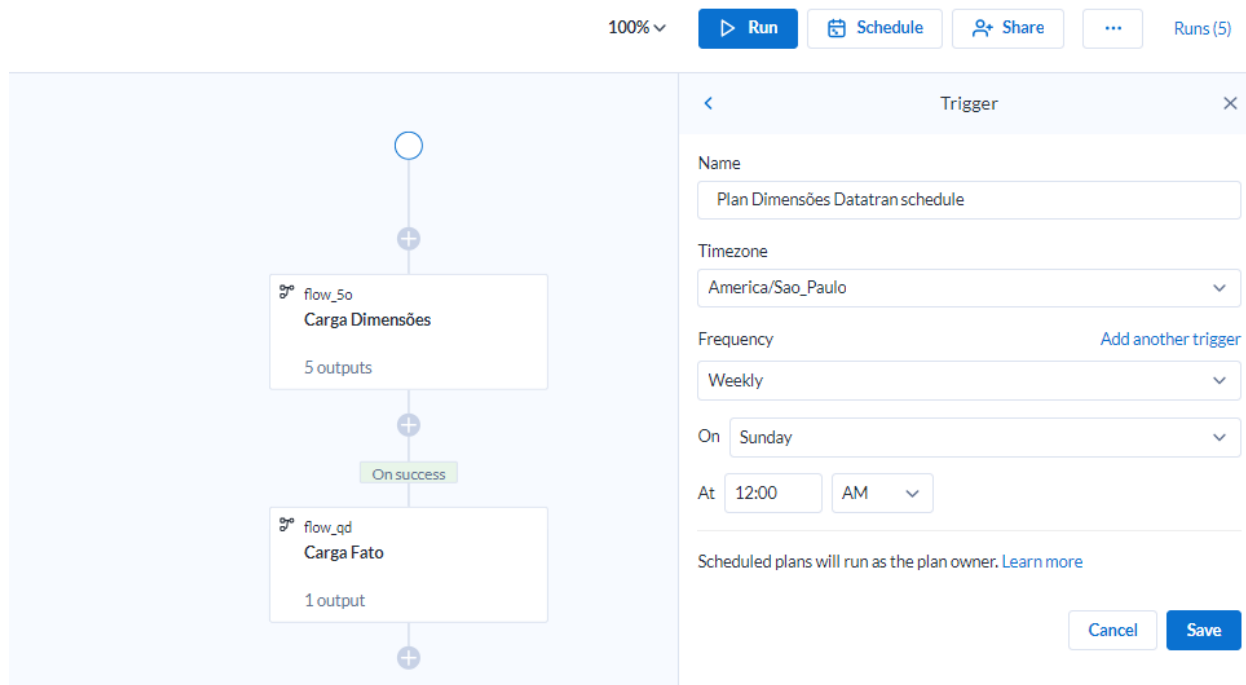
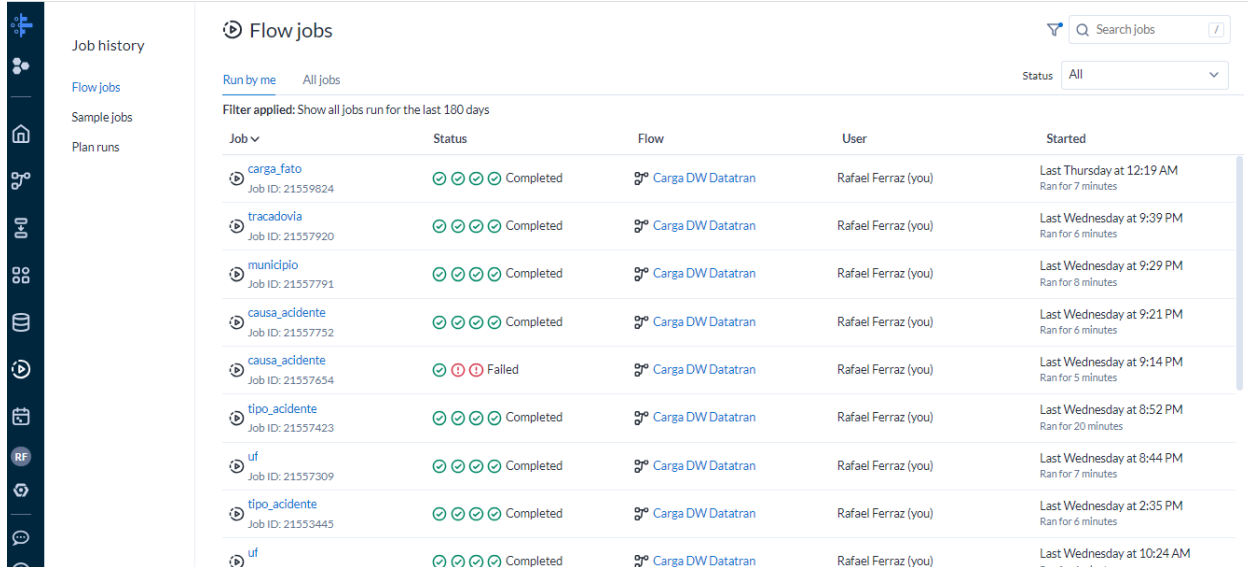


Figura 23 -Interface de agendamento de execuções do módulo Dataprep

Também é possível acompanhar o histórico das execuções (tanto manuais quando agendadas) a partir do módulo “Job History” da ferramenta:



Job	Status	Flow	User	Started
carga_fato Job ID: 21559824	Completed	Carga DW Datatran	Rafael Ferraz (you)	Last Thursday at 12:19 AM Ran for 7 minutes
tracadovia Job ID: 21557920	Completed	Carga DW Datatran	Rafael Ferraz (you)	Last Wednesday at 9:39 PM Ran for 6 minutes
município Job ID: 21557791	Completed	Carga DW Datatran	Rafael Ferraz (you)	Last Wednesday at 9:29 PM Ran for 8 minutes
causa_acidente Job ID: 21557752	Completed	Carga DW Datatran	Rafael Ferraz (you)	Last Wednesday at 9:21 PM Ran for 6 minutes
causa_acidente Job ID: 21557654	Failed	Carga DW Datatran	Rafael Ferraz (you)	Last Wednesday at 9:14 PM Ran for 5 minutes
tipo_acidente Job ID: 21557423	Completed	Carga DW Datatran	Rafael Ferraz (you)	Last Wednesday at 8:52 PM Ran for 20 minutes
uf Job ID: 21557309	Completed	Carga DW Datatran	Rafael Ferraz (you)	Last Wednesday at 8:44 PM Ran for 7 minutes
tipo_acidente Job ID: 21553445	Completed	Carga DW Datatran	Rafael Ferraz (you)	Last Wednesday at 2:35 PM Ran for 6 minutes
uf	Completed	Carga DW Datatran	Rafael Ferraz (you)	Last Wednesday at 10:24 AM

Figura 24 - Histórico de execuções do Dataprep

10. Análise

Foi executada uma análise sobre os dados submetidos ao pipeline no esquema de dados relacional para avaliação da qualidade das informações, buscando assim resultados satisfatórios para a resolução do problema proposto neste MVP.

10.1. Qualidade dos Dados

Para análise da qualidade de dados, utilizei instruções em SQL diretamente nos campos das tabelas do modelo de dados relacional para constatar os tipos de dados e domínios das informações, de forma que, se a contagem for maior do que 0 (zero), os dados da tabela possuem inconsistências.

Tabela: dm_causaacidente

Campo	Regra(s)	Código SQL da Análise	Saída	Resultado
id_causaacidente	- Inteiro - Não Nulo	SELECT count(*) FROM dm_causaacidente WHERE id_causaacidente NOT REGEXP '^-[0-9]+\$' OR id_causaacidente IS NULL;	0	Satisfatório
ds_name	- Não Nulo	SELECT count(*)FROM dm_causaacidenteWHERE TRIM(ds_name) IS NULL	0	Satisfatório
dt_alteracao	- Data/Hora Válidos	SELECT count(*) FROM dm_causaacidente WHERE DAYNAME(dt_alteracao) IS NULL;	0	Satisfatório

Tabela 7 - Análise de qualidade da Tabela dm_causaacidente

Tabela: dm_municipio

Campo	Regra(s)	Código SQL da Análise	Saída	Resultado
id_municipio	- Inteiro - Não Nulo	SELECT count(*) FROM dm_municipio WHERE id_municipio NOT REGEXP '^- ?[0-9]+\$' OR id_municipio IS NULL;	0	Satisfatório
ds_name	- Não Nulo	SELECT count(*) FROM dm_municipio WHERE TRIM(ds_name) IS NULL	0	Satisfatório
dt_alteracao	- Data/Hora Válidos	SELECT count(*) FROM dm_municipio WHERE DAYNAME(dt_alteracao) IS NULL;	0	Satisfatório

Tabela 8 - Análise de qualidade da Tabela dm_municipio

Tabela: dm_tipoacidente

Campo	Regra(s)	Código SQL da Análise	Saída	Resultado
id_tipoacidente	- Inteiro - Não Nulo	SELECT count(*) FROM dm_tipoacidente WHERE id_tipoacidente NOT REGEXP '^-[0-9]+\$' OR id_tipoacidente IS NULL;	0	Satisfatório
ds_name	- Não Nulo	SELECT count(*) FROM dm_tipoacidente WHERE TRIM(ds_name) IS NULL	0	Satisfatório
dt_alteracao	- Data/Hora Válidos	SELECT count(*) FROM dm_tipoacidente WHERE DAYNAME(dt_alteracao) IS NULL;	0	Satisfatório

Tabela 9 - Análise de qualidade da Tabela dm_tipoacidente

Tabela: dm_tracadovia

Campo	Regra(s)	Código SQL da Análise	Saída	Resultado
id_tracadovia	- Inteiro - Não Nulo	SELECT count(*) FROM dm_tracadovia WHERE id_tracadovia NOT REGEXP '^- ?[0-9]+\$' OR id_tracadovia IS NULL;	0	Satisfatório
ds_name	- Não Nulo	SELECT count(*) FROM dm_tracadovia WHERE TRIM(ds_name) IS NULL	0	Satisfatório
dt_alteracao	- Data/Hora Válidos	SELECT count(*) FROM dm_tracadovia WHERE DAYNAME(dt_alteracao) IS NULL;	0	Satisfatório

Tabela 10 - Análise de qualidade da Tabela dm_tracadovia

Tabela: dm_uf

Campo	Regra(s)	Código SQL da Análise	Saída	Resultado
id_uf	- Inteiro - Não Nulo	SELECT count(*) FROM dm_uf WHERE id_uf NOT REGEXP '^-[0-9]+\$' OR id_uf IS NULL;	0	Satisfatório
ds_name	- Não Nulo	SELECT count(*) FROM dm_uf WHERE TRIM(ds_name) IS NULL	0	Satisfatório
dt_alteracao	- Data/Hora Válidos	SELECT count(*) FROM dm_uf WHERE DAYNAME(dt_alteracao) IS NULL;	0	Satisfatório

Tabela 11 - Análise de qualidade da Tabela dm_uf



DOCUMENTAÇÃO DO MVP
Sprint: Engenharia de Dados

Tabela: ft_ocorrencias

Campo	Regra(s)	Código SQL da Análise	Saída	Resultado
id_ocorrencias	- Inteiro - Não Nulo	SELECT count(*) FROM ft_ocorrencias WHERE id_ocorrencias NOT REGEXP '^-[0-9]+\$' OR id_ocorrencias IS NULL;	0	Satisfatório
id_uf	- Inteiro - Não Nulo	SELECT count(*) FROM ft_ocorrencias WHERE id_uf NOT REGEXP '^-[0-9]+\$' OR id_uf IS NULL;	0	Satisfatório
id_tracadovia	- Inteiro - Não Nulo	SELECT count(*) FROM ft_ocorrencias WHERE id_tracadovia NOT REGEXP '^-[0-9]+\$' OR id_tracadovia IS NULL;	0	Satisfatório
id_causaacidente	- Inteiro - Não Nulo	SELECT count(*) FROM ft_ocorrencias WHERE id_causaacidente NOT REGEXP '^-[0-9]+\$' OR id_causaacidente IS NULL;	0	Satisfatório
id_municipio	- Inteiro - Não Nulo	SELECT count(*) FROM ft_ocorrencias WHERE id_municipio NOT REGEXP '^-[0-9]+\$' OR id_municipio IS NULL;	0	Satisfatório
id_tipoacidente	- Inteiro - Não Nulo	SELECT count(*) FROM ft_ocorrencias WHERE id_tipoacidente NOT REGEXP '^-[0-9]+\$' OR id_tipoacidente IS NULL;	0	Satisfatório
dt_dataocorrencia	- Data Válida - Não Nulo	SELECT count(*) FROM ft_ocorrencias WHERE DAYNAME(dt_dataocorrencia) IS NULL OR dt_dataocorrencia IS NULL;	0	Satisfatório

Campo	Regra(s)	Código SQL da Análise	Saída	Resultado
ds_diasemana	- [segunda-feira, terça-feira, quarta-feira, quinta-feira, sexta-feira, sábado, domingo] - Não Nulo	SELECT count(*) FROM ft_ocorrencias WHERE ds_diasemana NOT IN ('segunda-feira', 'terça-feira', 'quarta-feira', 'quinta-feira', 'sexta-feira', 'sábado', 'domingo') OR TRIM(ds_diasemana) IS NULL	0	Satisfatório
ds_horario	- Hora Válida	SELECT count(*) FROM ft_ocorrencias WHERE STR_TO_DATE(ds_horario, '%Y-%m-%d %H:%i:%s') IS NULL	0	Satisfatório
int_br	- Inteiro	SELECT count(*) FROM ft_ocorrencias WHERE int_br NOT REGEXP '^-[0-9]+\$'	0	Satisfatório
ds_km	- Numérico - [NA]	SELECT count(*) FROM ft_ocorrencias WHERE ds_km NOT REGEXP '^[0-9]+\.\?[0-9]*\$' AND UPPER(ds_km) != 'NA';	0	Satisfatório
ds_classificacao	- [Com Vítimas Feridas, Com Vítimas Fatais, Sem Vítimas] - Não Nulo	SELECT count(*) FROM ft_ocorrencias WHERE ds_classificacao NOT IN ('Com Vítimas Feridas', 'Com Vítimas Fatais', 'Sem Vítimas')	0	Satisfatório
ds_fasedia	- [Pleno dia, Plena Noite, Amanhecer, Anoitecer]	SELECT count(*) FROM ft_ocorrencias WHERE ds_fasedia NOT IN ('Pleno dia', 'Plena Noite', 'Amanhecer', 'Anoitecer')	0	Satisfatório
ds_sentidovia	- [Crescente, Decrescente, Não Informado]	SELECT count(*) FROM ft_ocorrencias WHERE ds_sentidovia NOT IN ('Crescente', 'Decrescente', 'Não Informado')	0	Satisfatório

Campo	Regra(s)	Código SQL da Análise	Saída	Resultado
ds_condicaometereologica		SELECT count(*) FROM ft_ocorrencias WHERE ds_condicaometereologica NOT IN ('Céu Claro', 'Nublado', 'Chuva', 'Garoa/Chuvisco', 'Sol', 'Ignorado', 'Vento', 'Nevoeiro/Neblina', 'Granizo', 'Neve')	0	Satisfatório
ds_tipopista	- [Dupla, Simples, Múltipla]	SELECT count(*) FROM ft_ocorrencias WHERE ds_tipopista NOT IN ('Dupla', 'Simples', 'Múltipla')	0	Satisfatório
int_pessoas	- Inteiro	SELECT count(*) FROM ft_ocorrencias WHERE int_pessoas NOT REGEXP '^-[0-9]+\$' OR int_pessoas IS NULL;	0	Satisfatório
int_mortos	- Inteiro	SELECT count(*) FROM ft_ocorrencias WHERE int_mortos NOT REGEXP '^-[0-9]+\$' OR int_mortos IS NULL;	0	Satisfatório
int_feridosleves	- Inteiro	SELECT count(*) FROM ft_ocorrencias WHERE int_feridosleves NOT REGEXP '^-[0-9]+\$' OR int_feridosleves IS NULL;	0	Satisfatório
int_feridosgraves	- Inteiro	SELECT count(*) FROM ft_ocorrencias WHERE int_feridosgraves NOT REGEXP '^-[0-9]+\$' OR int_feridosgraves IS NULL;	0	Satisfatório
int_ilesos	- Inteiro	SELECT count(*) FROM ft_ocorrencias WHERE int_ilesos NOT REGEXP '^-[0-9]+\$' OR int_ilesos IS NULL;	0	Satisfatório

Campo	Regra(s)	Código SQL da Análise	Saída	Resultado
int_ignorados	- Inteiro	SELECT count(*) FROM ft_ocorrencias WHERE int_ignorados NOT REGEXP '^-[0-9]+\$' OR int_ignorados IS NULL;	0	Satisfatório
int_feridos	- Inteiro	SELECT count(*) FROM ft_ocorrencias WHERE int_feridos NOT REGEXP '^-[0-9]+\$' OR int_feridos IS NULL;	0	Satisfatório
int_veiculos	- Inteiro	SELECT count(*) FROM ft_ocorrencias WHERE int_veiculos NOT REGEXP '^-[0-9]+\$' OR int_veiculos IS NULL;	0	Satisfatório
ds_latitude	- Separador: . - Valor: entre -90 e 90	SELECT count(*) FROM ft_ocorrencias WHERE (cast(ds_latitude as decimal(10,8)) < -90 and cast(ds_latitude as decimal(10,8)) > 90) OR ds_latitude LIKE '%,%'	173.068	Insatisfatório. Para correção seria necessário substituir todos os caracteres de “,” por “.”, de forma a padronizar com o formato universal.
ds_longitude	- Separador: . - Valor: entre -180 e 180	SELECT count(*) FROM ft_ocorrencias WHERE (cast(ds_longitude as decimal(10,8)) < -180 and cast(ds_longitude as decimal(10,8)) > 180) OR ds_longitude LIKE '%,%'	173.067	Insatisfatório. Para correção seria necessário substituir todos os caracteres de “,” por “.”, de forma a padronizar com o formato universal.
dt_alteracao	- Data/Hora Válidos	SELECT count(*) FROM ft_ocorrencias WHERE DAYNAME(dt_alteracao) IS NULL;	0	Satisfatório

10.2. Solução do Problema

Após todo o processo de coleta, tratamento, armazenamento e validação de qualidade dos dados, chega o momento de responder aos questionamentos elaborados no começo deste trabalho.

Assim como para a validação de qualidade dos dados, utilizei a linguagem SQL para responder aos questionamentos.

10.2.1. Perguntas

Pergunta 1: Qual o evolutivo de acidentes por ano?																
Descrição:	Agrupar a quantidade de acidentes ocorridos por ano, considerando os últimos 10 anos ou anos disponíveis nas bases de dados da PRF.															
Código SQL	SELECT year(dt_dataocorrencia) as 'Ano' ,count(*) as 'Qtd. Acidentes' FROM ft_ocorrencias GROUP BY year(dt_dataocorrencia) ORDER BY year(dt_dataocorrencia)															
Resultado do Código	<table><tr><td></td><td>🔒 123 Ano ▼</td><td>123 Qtd. Acidentes ▼</td></tr><tr><td>1</td><td>2020</td><td>63576</td></tr><tr><td>2</td><td>2021</td><td>64539</td></tr><tr><td>3</td><td>2022</td><td>64547</td></tr><tr><td>4</td><td>2023</td><td>44035</td></tr></table>		🔒 123 Ano ▼	123 Qtd. Acidentes ▼	1	2020	63576	2	2021	64539	3	2022	64547	4	2023	44035
	🔒 123 Ano ▼	123 Qtd. Acidentes ▼														
1	2020	63576														
2	2021	64539														
3	2022	64547														
4	2023	44035														
Discussão	A partir do resultado obtido, nota-se que existe uma média de 64.220 ocorrência de acidentes (considerando os anos completos de 2020 até 2022) por ano.															

Tabela 12 - Pergunta 1

Pergunta 2: Qual o evolutivo de vítimas em acidentes por ano e mês, considerando o quantitativo de pessoas envolvidas X fatalidades?	
Descrição:	Comparativo cronológico por ano/mês sobre a quantidade de pessoas envolvidas em acidentes, paralelo ao quantitativo de pessoas que vieram a óbito, considerando os últimos 5 anos ou anos disponíveis na base da PRF.
Código SQL	<pre>SELECT YEAR(dt_dataocorrencia) as "Ano" ,MONTH(dt_dataocorrencia) as "Mês" ,SUM(int_pessoas) as "Qtd. Pessoas" ,SUM(int_mortos) as "Qtd. Mortos" FROM ft_ocorrencias GROUP BY YEAR(dt_dataocorrencia) ,MONTH(dt_dataocorrencia) ORDER BY YEAR(dt_dataocorrencia) ,MONTH(dt_dataocorrencia)</pre>

Resultado do Código	123 Ano	123 Mês	123 Qtd. Pessoas	123 Qtd. Mortos
1	2020	1	14070	410
2	2020	2	13192	386
3	2020	3	10777	391
4	2020	4	8385	355
5	2020	5	10080	387
6	2020	6	10453	359
7	2020	7	11633	455
8	2020	8	12884	455
9	2020	9	13205	499
10	2020	10	14515	488
11	2020	11	13496	526
12	2020	12	15054	582
13	2021	1	13203	472
14	2021	2	11431	348
15	2021	3	10663	400
16	2021	4	11003	379
17	2021	5	12727	455
18	2021	6	12029	422
19	2021	7	13753	521
20	2021	8	12891	453
21	2021	9	12855	465
22	2021	10	13625	510
23	2021	11	12180	426
24	2021	12	14565	545
25	2022	1	12507	453
26	2022	2	11399	397
27	2022	3	12248	439
28	2022	4	12671	393
29	2022	5	12670	491
30	2022	6	11817	405
31	2022	7	14078	494
32	2022	8	13487	486
33	2022	9	13336	498
34	2022	10	14039	496
35	2022	11	12428	417
36	2022	12	14586	470
37	2023	1	13751	448
38	2023	2	11639	334
39	2023	3	13151	464
40	2023	4	13628	403
41	2023	5	13686	475
42	2023	6	14157	546
43	2023	7	15174	557
44	2023	8	13229	484

Discussão	A partir do resultado é possível perceber que a quantidade de pessoas envolvidas em acidentes aumenta nos últimos meses de cada ano. Ordenando a pesquisa por quantidade de mortos (forma decrescente) é possível perceber que os maiores índices também estão relacionados aos últimos meses do ano, a partir do segundo semestre.
------------------	---

Tabela 13 - Pergunta 2

Pergunta 3: Qual o percentual de pessoas envolvidas em acidentes por gênero?	
Descrição:	Percentual do total de vítimas em acidentes de toda a base coletada por gênero masculino, feminino e outros (sem informação de gênero na base).
Código SQL	N/A
Resultado do Código	Não foi possível estimar
Discussão	<p>As bases utilizadas neste trabalho extraídas do Portal de Dados Abertos da PRF foram aquelas agrupadas por ocorrência (Documento CSV de Acidentes Agrupados por ocorrência), não trazendo informações individualizadas de pessoas envolvidas no acidente, apenas os totalizadores.</p> <p>A motivação por utilizar bases reduzidas (agrupadas por ocorrência) fornecidas pela PRF foram os desafios encontrados ao processar estes dados no ETL, devido ao baixo nível computacional das instâncias do Google Cloud, visando economia de créditos gratuitos.</p>

Tabela 14 - Pergunta 3

Pergunta 4: Quais são as “Top 10” causas de acidente?	
Descrição:	10 maiores causas de acidentes considerando toda a base coletada.
Código SQL	<pre>SELECT count(*) AS "Qtd. Acidentes" ,dc.ds_name AS "Causa Acidente" FROM ft_ocorrencias fo INNER JOIN dm_causaacidente dc ON (fo.id_causaacidente=dc.id_causaacidente) GROUP BY dc.ds_name ORDER BY count(*) DESC LIMIT 10</pre>

Resultado do Código	123 Qtd. Acidentes ▾ ABC Causa Acidente ▾	
	1	22407 Falta de Atenção à Condução
	2	21728 Reação tardia ou ineficiente do condutor
	3	20408 Velocidade Incompatível
	4	19157 Ausência de reação do condutor
	5	14974 Acessar a via sem observar a presença dos outros veículos
	6	12263 Condutor deixou de manter distância do veículo da frente
	7	11227 Ingestão de álcool pelo condutor
	8	9908 Manobra de mudança de faixa
	9	8800 Desobediência às normas de trânsito pelo condutor
	10	7906 Condutor Dormindo
Discussão	A partir da observação das causas que compõem o topo da lista, é possível concluir que as características mais presentes estão relacionadas à falta de atenção dos condutores, imprudência e imperícia.	

Tabela 15 - Pergunta 4

Pergunta 5:	Quais são os “Top 20” modelos de veículos envolvidos em acidentes?
Descrição:	20 veículos mais envolvidos em acidentes considerando toda a base coletada.
Código SQL	N/A
Resultado do Código	Não foi possível estimar
Discussão	Assim como na questão 3, também não foi possível estimar com este nível de detalhes, visto que os datasets disponibilizados pela PRF contendo informações individualizadas por pessoas e veículos são grandes para as capacidades de máquina configuradas para este trabalho, o que impacta diretamente na performance nas etapas mais críticas do pipeline.

Tabela 16 - Pergunta 5

Pergunta 6:	Qual o quantitativo de acidentes por dia da semana?
Descrição:	Quantitativo de acidentes por dia da semana (Domingo, Segunda-feira, (...), Sexta-feira) considerando toda a base coletada.
Código SQL	<pre>SELECT ds_diasemana as "Dia da Semana" ,count(*) as "Qtd. Acidentes" FROM ft_ocorrencias GROUP BY ds_diasemana ORDER BY count(*) DESC</pre>

Resultado do Código	ABC Dia da Semana	123 Qtd. Acidentes	123 Total Pessoas Envolvidas
	1 sábado	39591	94571
	2 domingo	39574	96474
	3 sexta-feira	36547	87583
	4 segunda-feira	32063	75922
	5 quinta-feira	30544	71175
	6 quarta-feira	29635	69387
	7 terça-feira	28743	67238
Discussão	<p>A partir da observação do resultado acima, é possível constatar que a maioria dos acidentes ocorrem, consecutivamente, nos dias de sábado e domingo, diferente do que eu imaginava anteriormente de que eles ocorriam nos primeiros dias úteis da semana.</p> <p>Adicionalmente e fora do escopo da pergunta, incluí a coluna do total de pessoas envolvidas, constatando assim que apesar do dia com maior número de ocorrências, é o segundo maior dia (domingo) que possui o maior número de pessoas envolvidas no acidente. Ou seja, veículos com um número ligeiramente maior de pessoas.</p>		

Tabela 17 - Pergunta 6

Pergunta 7: Qual o quantitativo de acidentes por faixa de horário?	
Descrição:	Comparativo de acidentes ocorridos por faixa de horário (Exemplo: 13 = 13:00 até 13:59), considerando toda a base coletada.
Código SQL	<pre>SELECT lpad(hour(ds_horario),2,'0') AS "Faixa Horário", count(*) as "Qtd. Acidentes" FROM ft_ocorrencias WHERE ds_diasemana != 'domingo' GROUP BY lpad(hour(ds_horario),2,'0') ORDER BY lpad(hour(ds_horario),2,'0')</pre>

Resultado do Código	ABC Faixa Horário ▾ 123 Qtd. Acidentes ▾	
	1	00 3988
	2	01 3337
	3	02 2998
	4	03 3076
	5	04 3856
	6	05 5586
	7	06 8756
	8	07 12672
	9	08 10361
	10	09 8402
	11	10 8335
	12	11 8930
	13	12 8416
	14	13 8714
	15	14 9625
	16	15 10218
	17	16 11000
	18	17 13175
	19	18 14797
	20	19 12455
	21	20 9113
	22	21 7397
	23	22 6496
	24	23 5420
Discussão	<p>Aqui é possível observar um padrão de crescimento no número de acidentes às 6, 7 e 8 da manhã e outro pico nos horários de 17, 18 e 19 da tarde, que são normalmente os horários com maior movimentação de veículos na rua dado o expediente de trabalho.</p> <p>Adicionalmente e fora do escopo da questão, fiz uma condição para ignorar o dia de domingo e focar os resultados com ocorrências de acidentes nos dias úteis.</p>	

Tabela 18 - Pergunta 7

Pergunta 8:	Qual o quantitativo de acidentes por faixa etária?
Descrição:	Volumetria de vítimas em acidentes por faixa etária (Exemplo: 0 a 5, 6 a 10, 11 a 15, 16 a 20, 21 a 25, (...), 80 e mais)
Código SQL	N/A
Resultado do Código	Não foi possível estimar

Discussão	Não foi possível estimar com este nível de detalhes, visto que os datasets disponibilizados pela PRF contendo informações individualizadas por pessoas e veículos são grandes para as capacidades de máquina configuradas para este trabalho, o que impacta diretamente na performance nas etapas mais críticas do pipeline.
------------------	--

Tabela 19 - Pergunta 8

Pergunta 9:	Qual o quantitativo de vítimas X fatalidades em acidentes por Estado?
Descrição:	Comparativo do quantitativo de vítimas e vítimas fatais por Estado.
Código SQL	<pre>SELECT du.ds_name as "Estado" , sum(int_pessoas) as "Pessoas Envolvidas" , sum(int_mortos) as "Vítimas Fatais" FROM ft_ocorrencias fo INNER JOIN dm_uf du ON (fo.id_uf=du.id_uf) GROUP BY du.ds_name ORDER BY sum(int_pessoas) DESC</pre>

Resultado do Código	ABC Estado	123 Pessoas Envolvidas	123 Vítimas Fatais
	MG	73737	2588
	SC	63577	1337
	PR	62170	2039
	RS	41332	1108
	RJ	39921	1051
	SP	36738	834
	BA	31880	1899
	GO	28143	1043
	PE	23195	1123
	ES	20453	558
	MT	18889	901
	MS	14128	573
	RO	12749	322
	CE	12576	598
	PB	12260	425
	RN	10923	356
	MA	10640	824
	PI	9908	523
	DF	8682	155
	PA	8535	593
	TO	5402	306
	AL	4991	270
	SE	4548	153
	AC	2170	69
	RR	2130	95
	AP	1520	46
	AM	1153	50
Discussão	<p>Com base nos resultados é possível verificar que o maior envolvimento de pessoas em acidentes ocorre no estado de Minas Gerais e consequentemente, o maior número de fatalidades, diferente do senso comum em que se imagina que a maioria dos números ocorriam estados como SP e RJ, dada a densidade de pessoas nas metrópoles.</p> <p>Outro ponto interessante é que em alguns estados a quantidade de óbitos não é proporcional à quantidade de pessoas envolvidas, como no estado de São Paulo, aonde apesar do número alto de pessoas envolvidas, os óbitos são menores em relação à média e no estado do Paraná, onde o número de óbitos é maior em relação aos dos estados vizinhos neste mesmo indicador.</p>		

Tabela 20 - Pergunta 9

Pergunta 10: Quais são as 10 maiores causas de fatalidades nos acidentes?																																			
Descrição:	As 10 maiores causas de acidentes considerando a quantidade de vítimas que vieram a óbito.																																		
Código SQL	<pre>SELECT dt.ds_name as "Tipo Acidente" ,sum(int_mortos) as "Qtd. Mortos" FROM ft_ocorrencias fo INNER JOIN dm_tipoacidente dt on (fo.id_tipoacidente=dt.id_tipoacidente) GROUP BY dt.ds_name ORDER BY sum(int_mortos) DESC LIMIT 10</pre>																																		
Resultado do Código	<table border="1"> <thead> <tr> <th></th><th>ABC Tipo Acidente</th><th>123 Qtd. Mortos</th></tr> </thead> <tbody> <tr><td>1</td><td>Colisão frontal</td><td>6153</td></tr> <tr><td>2</td><td>Atropelamento de Pedestre</td><td>3201</td></tr> <tr><td>3</td><td>Saída de leito carroçável</td><td>2482</td></tr> <tr><td>4</td><td>Colisão traseira</td><td>2087</td></tr> <tr><td>5</td><td>Colisão transversal</td><td>1551</td></tr> <tr><td>6</td><td>Tombamento</td><td>950</td></tr> <tr><td>7</td><td>Colisão com objeto</td><td>757</td></tr> <tr><td>8</td><td>Colisão lateral mesmo sentido</td><td>519</td></tr> <tr><td>9</td><td>Colisão lateral sentido oposto</td><td>511</td></tr> <tr><td>10</td><td>Queda de ocupante de veículo</td><td>326</td></tr> </tbody> </table>			ABC Tipo Acidente	123 Qtd. Mortos	1	Colisão frontal	6153	2	Atropelamento de Pedestre	3201	3	Saída de leito carroçável	2482	4	Colisão traseira	2087	5	Colisão transversal	1551	6	Tombamento	950	7	Colisão com objeto	757	8	Colisão lateral mesmo sentido	519	9	Colisão lateral sentido oposto	511	10	Queda de ocupante de veículo	326
	ABC Tipo Acidente	123 Qtd. Mortos																																	
1	Colisão frontal	6153																																	
2	Atropelamento de Pedestre	3201																																	
3	Saída de leito carroçável	2482																																	
4	Colisão traseira	2087																																	
5	Colisão transversal	1551																																	
6	Tombamento	950																																	
7	Colisão com objeto	757																																	
8	Colisão lateral mesmo sentido	519																																	
9	Colisão lateral sentido oposto	511																																	
10	Queda de ocupante de veículo	326																																	
Discussão	<p>Nessa pergunta tomei a liberdade de modificar a “causa” do acidente por “tipo” do acidente, dada a similaridade com a pergunta 4 e a oportunidade de abordar dados coletados que não foram utilizados.</p> <p>Podemos observar por este resultado que o tipo do acidente que ocupa o maior ranking da lista é exponencialmente maior que o segundo lugar e os demais, indicando um alarmante número de vidas perdidas em colisões frontais.</p>																																		

Tabela 21 - Pergunta 10

10.2.2. Discussão Geral Sobre as Perguntas

Entre as perguntas não respondidas, a principal causa se dá pela ausência das informações nos datasets coletados no portal de dados da PRF que, conforme citado anteriormente, optei por utilizar bases menores, agrupadas por ocorrência o que configura a ausência de informações de indivíduos envolvidos nas ocorrências. O portal atualmente disponibiliza uma maior riqueza de detalhes em outros datasets, como aqueles agrupados por pessoas.

11. Auto avaliação

Minha percepção foi muito positiva em relação à resolução do problema através das respostas aos questionamentos, sendo que apenas 3 das 10 perguntas não foram respondidas, possibilitando assim construir uma visão (mesmo que limitada) do panorama de acidentes nas rodovias federais.

No decorrer do desenvolvimento deste trabalho e exploração do ecossistema de nuvem do Google Cloud, elenquei uma série de melhorias que enriqueceriam grandemente o projeto.

No que diz respeito a performance e otimização, um aumento nos recursos computacionais das instâncias contratadas no Google Cloud possibilitaria a ingestão de datasets maiores e mais complexos na plataforma.

A omissão de algumas etapas poderia agilizar o processo, como a criação do modelo de dados em estrela diretamente no BigQuery. Ainda na ótica da omissão de etapas, a leitura dos dados brutos pela ferramenta de ETL poderia ser feita diretamente nos arquivos CSV, dispensando assim a necessidade da tabela Flat citada no capítulo 6.1.

Conforme citado anteriormente, na etapa de carga dos dados pela ferramenta DataPrep, eu optei em dado momento utilizar um script que “limpa” as tabelas do modelo de dados relacional antes de iniciar os procedimentos de carga, devido à dificuldade e esforço/tempo dedicados em solucionar o problema da integridade dos dados.

Nesta arquitetura de modelagem de dados (Dimensão e Fato) é altamente recomendado que a cada carga, os dados sejam atualizados ou acrescentados ao invés de apagados e carregados novamente a cada execução, visto que ferramentas importantes de visualização de dados e relatórios podem estar conectadas a estas fontes. Logo, uma melhoria para eliminação da solução de contorno que implementei contribuiria positivamente para o projeto.

Na etapa de coleta dos dados, o procedimento de acessar o Portal de Dados abertos, fazer o download dos arquivos e upload no Google Storage foi todo realizado de forma manual. A utilização de um script Python para fazer o “Scraping” automático e rotineiro destes dados seria de grande valia para a automação do projeto.

Outros pontos de melhoria que viriam a somar positivamente para o projeto:

- Integração com ferramenta de “dataviz” para mostrar o resultado das perguntas e outros indicadores a partir de painéis e gráficos. Isso também possibilitaria a utilização dos dados de coordenadas geográficas disponibilizados nos arquivos pela PRF para a criação de visões geográficas.
- Utilização de alguma ferramenta para gerenciamento do Catálogo de Dados, em substituição ao método manual que utilizei neste trabalho.
- Utilização da biblioteca Great Expectations (<https://greatexpectations.io/>) com Python para a análise de qualidade dos dados, em substituição ao método manual que utilizei neste trabalho.

- Integração com outras fontes de dados para aumentar a riqueza dos resultados obtidos. Ex.: API de feriados nacionais (<https://api.invertexto.com/api-feriados>), que possibilitaria coletar e deparar datas com a ocorrência dos acidentes em dias de feriado, aprofundando assim as visões sobre as principais causas nesses dias, locais com maior ocorrência ou o perfil do público envolvido nos acidentes.