Trabalho Prático - Armazém de Dados

Gabriela Tavares Barreto, Mirna Mendonça e Silva, Júlia Paes de Viterbo

Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte - MG - Brasil

1 Introdução

O Sistema de Relatório de Análise de Fatalidades (FARS) é um serviço público para coleta e análise de dados sobre acidentes fatais de veículos motorizados nos Estados Unidos. Criado com o objetivo de fornecer informações essenciais ao Estado para a elaboração e implementação de políticas públicas de segurança no trânsito, o FARS permite a interlocução direta entre autoridades, pesquisadores e a comunidade, promovendo a análise e compreensão aprofundada dos acidentes e suas características. O data warehouse do FARS foi desenvolvido para modelar esses dados de forma a possibilitar uma análise abrangente e detalhada, contribuindo para a identificação de tendências, aprimoramento da segurança viária e formulação de estratégias efetivas de prevenção de acidentes.

2 Processo de desenvolvimento do projeto

Pensamos inicialmente em escolher uma base de dados interessante e completa, de onde pudéssemos tirar conclusões relevantes, mas ao mesmo tempo que fosse simples e objetiva. Assim que encontramos uma base nesse padrão (de tema: Acidentes Veiculares ocorridos no EUA), caregados os dados com o uso do MySQL, fizemos um rascunho do esquema estrela e partimos para o tratamento da base escolhida com o PDI.

Importante apontar que fizemos um recorte na base para que fosse condizente com a capacidade armazenacional das máquinas a nosso dispor: selecionamos apenas os acidentes cujas informações fossem correpondentes ao ano de 2001. Após o tratamento citado, passamos à construção de um cubo Modrian para nosso armazém com o uso do Pentaho Workbench e, finalmente, utilizamos o Pentaho Report Designer para análises condizentes com os dados disponíveis, a exemplo de: nos estados presentes na base, analisar a distribuição das características dos acidentes, como forma de colisão, número de pedestres, número de motoristas bêbados etc.

3 Arquitetura

O Data Warehouse (DW) do Sistema de Relatório de Análise de Fatalidades (FARS) foi projetado para permitir uma análise abrangente e detalhada dos dados de acidentes fatais de veículos motorizados nos Estados Unidos. A arquitetura do DW é projetada com base em uma única tabela de fatos e duas dimensões principais: **tempo e lugar.**

Tabela de Fatos:

A tabela de fatos no DW do FARS contém informações detalhadas sobre cada acidente fatal

registrado. Ela inclui medidas quantitativas relevantes, como o número de fatalidades, a gravidade do acidente, informações sobre os veículos envolvidos, entre outros dados relevantes. Cada registro na tabela de fatos está associado a uma chave estrangeira para as dimensões de tempo e lugar.

Dimensão Tempo:

A dimensão de tempo é responsável por capturar informações relacionadas à data e hora do acidente fatal. Ela inclui atributos como ano, mês e dia. Esses atributos permitem a segmentação e análise dos dados com base no tempo, facilitando a identificação de tendências sazonais, e variações ao longo do tempo.

Dimensão Lugar: A dimensão de lugar abrange informações geográficas sobre os acidentes fatais. Ela inclui atributos como localização, estado, cidade, código postal, entre outros. Esses atributos permitem a análise espacial dos acidentes, identificando áreas de maior incidência, diferenças regionais e outras características geográficas relevantes.

A estrutura do DW permite uma modelagem dimensional que facilita consultas analíticas e exploração dos dados. Os usuários podem realizar análises complexas, como a contagem de acidentes por período de tempo, identificação de áreas com maior número de fatalidades, correlações entre fatores temporais e geográficos, entre outras análises relevantes para a segurança viária e formulação de estratégias de prevenção de acidentes.

O DW do FARS foi projetado levando em consideração a necessidade de uma análise abrangente dos dados, fornecendo insights valiosos para autoridades, pesquisadores e a comunidade em geral. Ao fornecer acesso fácil e estruturado aos dados de acidentes fatais, o DW contribui para a elaboração e implementação de políticas públicas de segurança no trânsito e o aprimoramento da segurança viária como um todo.

4 Descrição detalhada da modelagem dimensional

Para recordar, temos a modelagem de quatro passos composta por: 1. Selecionar o processo de negócios a modelar; 2. Declarar a granuliridade do processo; 3. Escolher as dimensões da tabela de fatos e 4. Identificar os fatos numéricos que irão populacionar a tabela-fato.

O primeiro passo foi completado assim que definimos nosso objeto de estudo como sendo os acidentes dos EUA em 2001. A base da FARS nos permitiu tal recorte.

Os passos do segundo ao quarto são fundamentados no esquema estrela mostrado anteriormente. Há duas tabelas de dimensão, cada uma com seus respectivos atributos e SKs correspondentes na tabela de fatos. Esses atributos foram retirados da base de dados original, em que todos com campos estavam juntos - aqui, separamos os correspondes a tempo e a lugar de forma a sermos conformes com os preceitos de DWs. Por fim, os fatos da tabela de fatos também foram selecionados dentre os campos da base primária, de modo que cada linha de entrada no DW contenha medidas relevantes para possíveis análises armazenadas nesta tabela.

5 Base de Dados

Com o objetivo de melhorar a segurança no trânsito, a Administração Nacional de Segurança no Trânsito em Rodovias (NHTSA, na sigla em inglês) criou o Sistema de Relatório de Análise de Fatalidades (FARS) em 1975. O banco de dados FARS contém informações sobre todos os acidentes fatais de veículos motorizados nos Estados Unidos.

Para ser incluído no FARS, um acidente deve envolver um veículo motorizado em uma via de trânsito normalmente aberta ao público e resultar na morte de uma pessoa (seja um ocupante de um veículo ou um pedestre) dentro de 30 dias do acidente. O arquivo FARS contém descrições de cada acidente fatal relatado. Cada caso possui mais de 100 elementos de dados codificados que caracterizam o acidente, os veículos e as pessoas envolvidas.

Os dados sobre acidentes de trânsito fatais envolvendo veículos motorizados são coletados a partir dos documentos de origem de cada estado e são codificados em formulários padrão do FARS. Os formulários do FARS geralmente incluem alguns ou todos os seguintes documentos: relatórios de acidentes policiais, arquivos de registro de veículos do estado, arquivos de habilitação de motoristas do estado, dados do departamento de rodovias do estado, estatísticas vitais, certificados de óbito, relatórios de médicos legistas/examinadores médicos, registros médicos hospitalares e relatórios de serviços médicos de emergência.

6 Acesso à Base de Dados

O acesso a base pode ser realizado mediante o link: Base de dados

7 ETL

Aqui descreve-se como foram criadas as dimensões de tempo e lugar e a tabela de fatos.

7.1 Extração

Para a dimensão de tempo, foram extraídas da tabela original as colunas *day* (dia), *month* (mês) e *year* (ano). Para a de lugar, as colunas *city* (cidade) e *state_nm* (nome do estado). As medidas numéricas da tabela de fatos foram extraídas das colunas *drunk_dr* (número de motoristas bêbados), *fatals* (fatalidades), *hit_run* (atropelamento com fuga), *c_m_zone* (indicador de zona de construção), *lgt_cond* (condição da iluminação), *peds* (quantidade de pedestres), *persons* (quantidade de pessoas), *ve_forms* (quantidade de veículos), *man_coll* (forma de colisão).

7.2 Transformação

Não foi necessária nenhuma transofrmação nas colunas da dimensão de tempo, que já estava bem tratada. A coluna de cidade precisou ser filtrada, já que alguns valores eram iguais a zero. Nas duas dimensões foram adicionadas sequências (de números inteiros) para serem as SKs. Na tabela de fatos, as colunas foram renomeadas e foi realizado o *join* com as SKs das dimensões.

7.3 Carga de dados

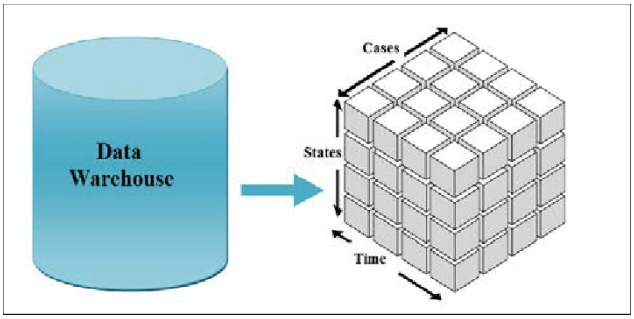
Cada dimensão e a fato foram carregadas para uma tabela específica no banco de dados que haviam sido criadas previamente usando SQL levando em conta os campos necessários para cada caso.

8 Outras ferramentas utilizadas e seus resultados

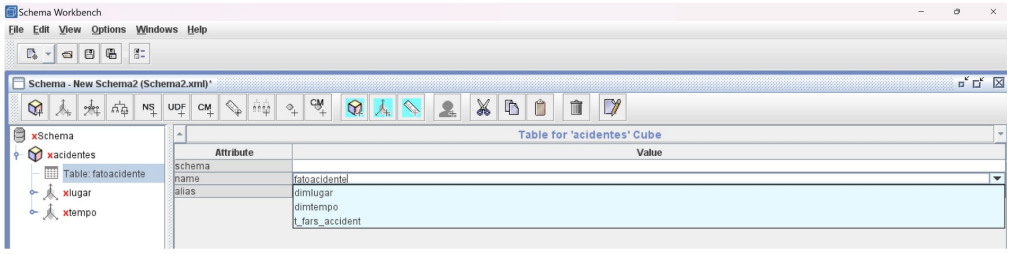
8.1 Pentaho Schema Workbench

O Pentaho Schema Workbench é uma das ferramentas do Pentaho utilizada para projetar modelos de dados e esquemas para análise multidimensional. Ela é uma interface gráfica que permite criar e gerenciar esquemas de cubos OLAP (Online Analytical Processing).

O Mondrian Cube é um mecanismo OLAP que capacita o Pentaho a compreender a modelagem multidimensional armazenada no banco de dados. Ele é composto por um arquivo XML que contém medidas numéricas, hierarquias e dimensões. Cada elemento do cubo faz referência a um campo específico de uma tabela no banco de dados.



Ao iniciar a criação do arquivo no Pentaho Schema Workbench, nós precisamos estabelecer uma conexão com a base de dados, bem como criar um novo esqueleto do cubo utilizando a funcionalidade "Add Cube". Nessa etapa, demos um nome para o cubo. Em seguida, associamos a tabela de fatos ao cubo e preencheemos o campo "Name"com o nome correspondente à tabela de fatos presente no banco de dados.



Após criar o esqueleto do cubo com a tabela de fatos associada, foi necessário estabelecer as

dimensões e suas respectivas hierarquias. Para cada dimensão atribuímos as respectivas "Foreign Key"como a chave estrangeira da tabela de fatos que se relaciona com essa dimensão.

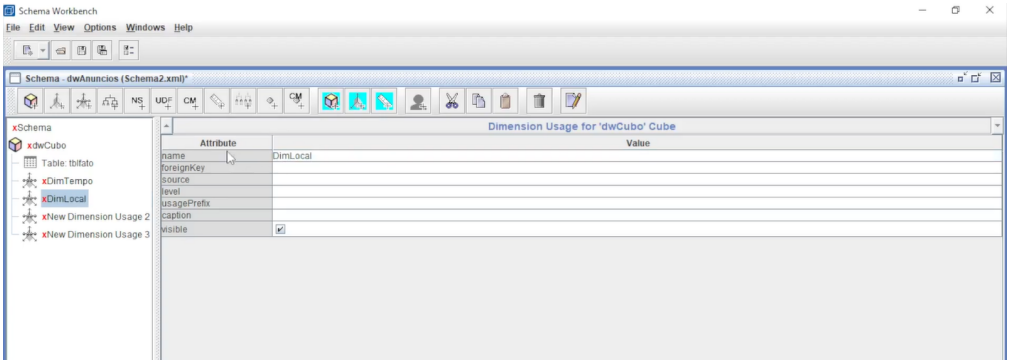


Figura 1: Dimensão local

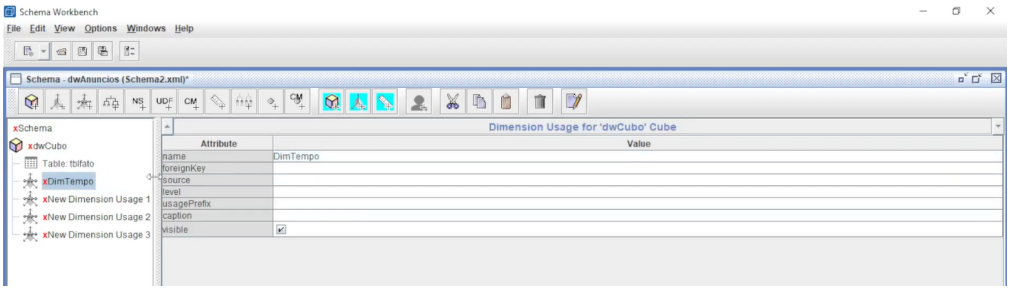


Figura 2: Dimensão tempo

Após criadas as dimensões, pudemos adicionar suas hierarquias e editar seus atributos de uma hierarquia. Para isso foi necessário vinculá-las à tabela de dimensão usando uma Primary Key.

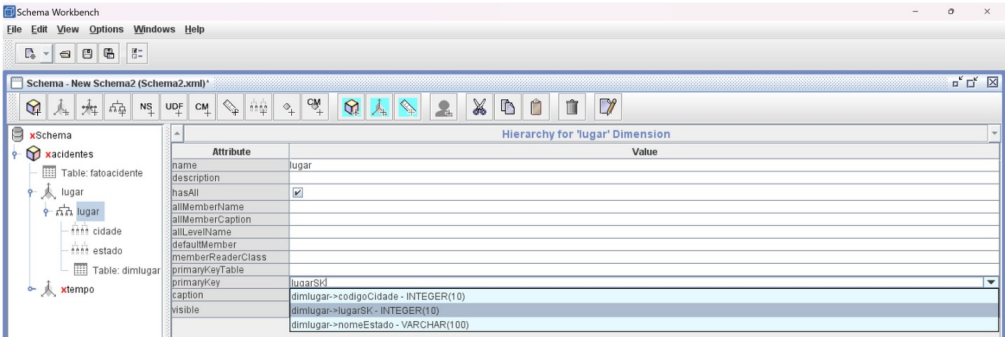


Figura 3: Criação de hierarquia

Com a hierarquia devidamente estabelecida,partimos para a fase de adicionar os níveis. Cada nível de uma hierarquia corresponde a um atributo da dimensão, ou seja, a uma coluna da tabela no banco de dados. Nos atentamos em adicionar os níveis em ordem de agregação na hierarquia, começando pelos níveis mais genéricos e finalizando com os mais específicos. Essa organização permite que o Pentaho Schema Workbench agregue corretamente os dados.

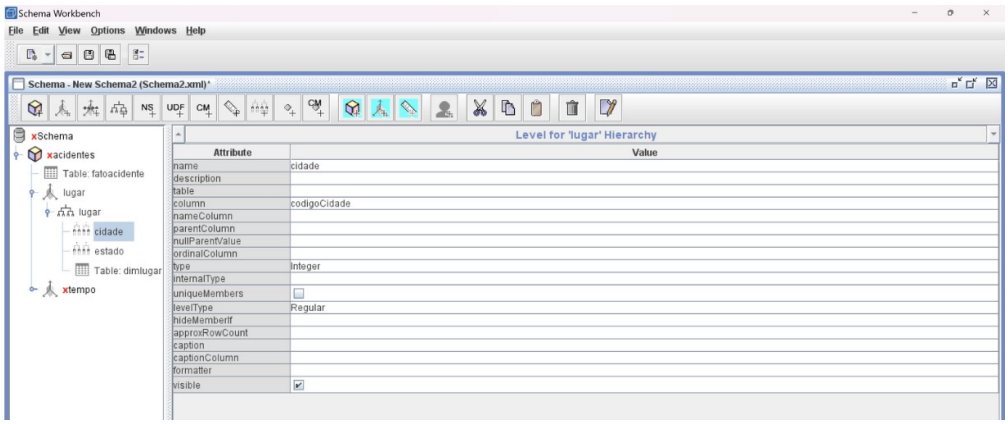


Figura 4: Adicionando os níveis

Depois disso adicionamos as medidas numéricas. Definimos que a medida seria calculada a partir de uma coluna numérica.

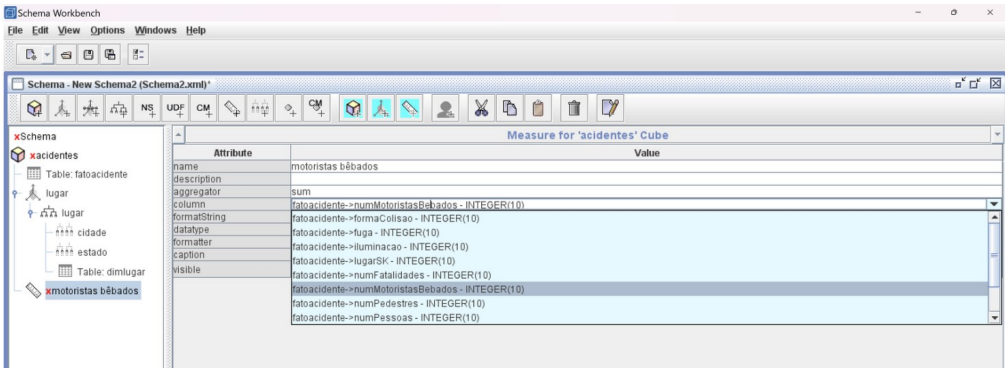


Figura 5: Adicionando a medida numérica

Por fim, publicamos o cubo no servidor Pentaho para utilizar o esquema como fonte de dados nas ferramentas de análise de negócios do Pentaho.

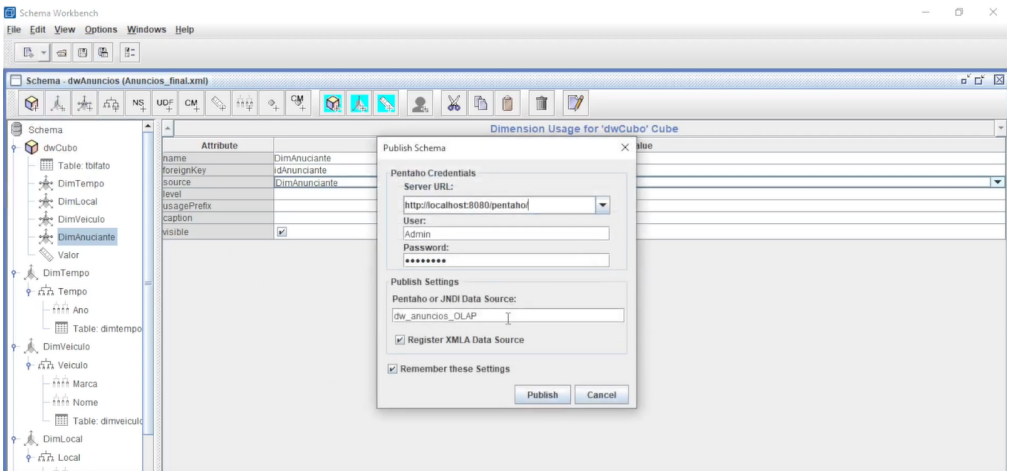


Figura 6: Publicando o servidor

8.2 Pentaho Report Designer

O Pentaho Report Designer foi uma ferramenta fundamental que utilizamos para analisar o Data Warehouse (DW) do FARS. Com essa poderosa plataforma, conseguimos explorar e extrair informações valiosas a partir dos dados disponíveis.

Ao iniciar nossa análise, utilizamos as funcionalidades do Pentaho Report Designer para selecionar as dimensões relevantes do DW, como localização geográfica, tipo de acidente, veículo envolvido e informações sobre as vítimas. Essas dimensões nos permitiram segmentar os dados e obter uma visão mais detalhada dos acidentes registrados no FARS.

Através das ferramentas de agregação e filtragem do Pentaho, conseguimos analisar estatísticas de acidentes, identificar tendências e padrões de ocorrência.

Com o Pentaho Report Designer, também conseguimos criar gráficos, tabelas e relatórios personalizados para visualizar e comunicar nossas descobertas. Essas visualizações foram essenciais para apresentar os dados de forma clara e compreensível.

Ao finalizar a análise do DW do FARS com o Pentaho Report Designer, obtivemos insights valiosos que contribuíram para o aprimoramento das políticas de segurança no trânsito. As informações extraídas nos permitiram compreender melhor os padrões de acidentes e suas causas, fornecendo subsídios para a implementação de medidas preventivas e ações que visam reduzir o número de fatalidades no trânsito.

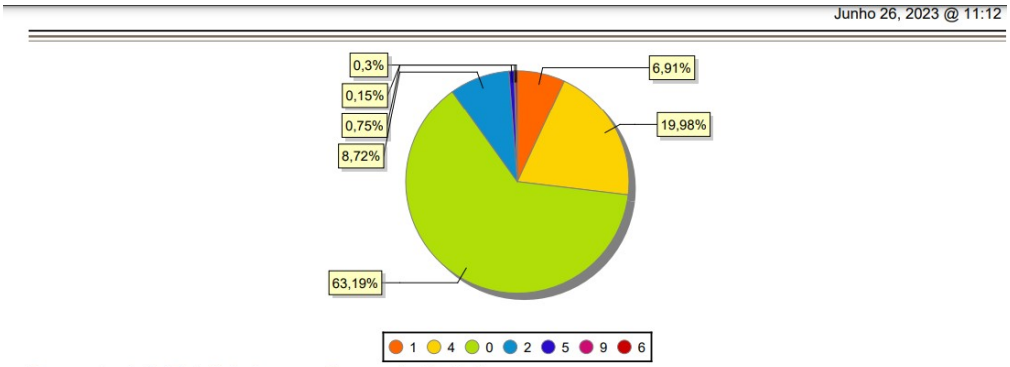
Em suma, o uso do Pentaho Report Designer foi fundamental para explorar e analisar o DW do FARS, permitindo-nos obter informações relevantes e contribuir para a melhoria da segurança viária. Sua interface intuitiva e recursos avançados nos proporcionaram uma análise abrangente e detalhada dos dados.

8.2.1 Relatórios obtidos

A partir do uso do Pentaho Report Designer, geramos relatórios que comparavam as medidas da tabela de fatos entre estados, cidades, e correlacionando isso com alguma segunda medida de referência, como o tipo de colisão, o dia do mês, a qualidade da iluminação etc.

O gráfico abaixo foi gerado em um desses relatórios. Ele apresenta a relação do número de

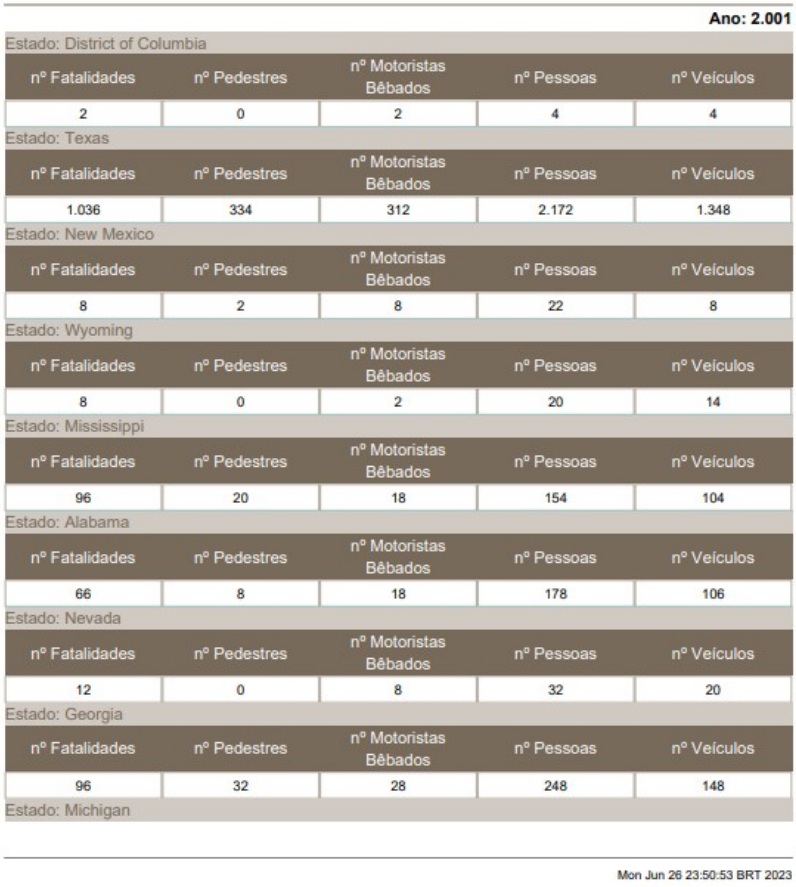
fatalidades de acordo com o tipo de forma de colisão.



Porcentual de Fatalidades por Forma de Colisão

0 Colisão não envolvendo veículo motorizado em transporte
1 Colisão traseira
2 Colisão frontal
3 Colisão traseira com traseira
4 Colisão em ângulo
5 Colisão superficial lateral: Mesma direção
6 Colisão superficial lateral: Direção oposta
9 Desconhecido

Segue um exemplo de relatório gerado, nesse caso, um que comparasse alguns dos fatos por estado.



Ano: 2.001

Estado: District of Columbia

nº Fatalidades	nº Pedestres	nº Motoristas Bêbados	nº Pessoas	nº Veículos
2	0	2	4	4

Estado: Texas

nº Fatalidades	nº Pedestres	nº Motoristas Bêbados	nº Pessoas	nº Veículos
1.036	334	312	2.172	1.348

Estado: New Mexico

nº Fatalidades	nº Pedestres	nº Motoristas Bêbados	nº Pessoas	nº Veículos
8	2	8	22	8

Estado: Wyoming

nº Fatalidades	nº Pedestres	nº Motoristas Bêbados	nº Pessoas	nº Veículos
8	0	2	20	14

Estado: Mississippi

nº Fatalidades	nº Pedestres	nº Motoristas Bêbados	nº Pessoas	nº Veículos
96	20	18	154	104

Estado: Alabama

nº Fatalidades	nº Pedestres	nº Motoristas Bêbados	nº Pessoas	nº Veículos
66	8	18	178	106

Estado: Nevada

nº Fatalidades	nº Pedestres	nº Motoristas Bêbados	nº Pessoas	nº Veículos
12	0	8	32	20

Estado: Georgia

nº Fatalidades	nº Pedestres	nº Motoristas Bêbados	nº Pessoas	nº Veículos
96	32	28	248	148

Estado: Michigan

Mon Jun 26 23:50:53 BRT 2023

9 Dificuldades encontradas no projeto

Durante o desenvolvimento do nosso projeto, enfrentamos algumas dificuldades que foram superadas com esforço e dedicação. Compartilhamos abaixo as principais dificuldades encontradas:

- **Acesso às bases:** Para acessar as bases de dados, enfrentamos desafios na configuração inicial da conexão com o banco de dados MySQL. Foi necessário compreender e aplicar corretamente as configurações de conexão, além de lidar com restrições e permissões de acesso impostas pelo banco de dados. Foi necessário realizar pesquisas, buscar suporte da comunidade e realizar ajustes até obtermos a conexão adequada.
- **Parte de extração e limpeza das bases (ETL):** Durante a etapa de extração e limpeza dos dados, nos deparamos com desafios na identificação e tratamento de dados ausentes, inconsistentes e duplicados nas bases de dados do FARS. Foi necessário utilizar o Pentaho Data Integration para implementar transformações e regras de limpeza dos dados, a fim de garantir a qualidade e integridade dos dados carregados no Data Warehouse. Lidar com grandes volumes de dados também exigiu otimização de desempenho e gerenciamento cuidadoso dos recursos disponíveis.
- **Utilização das ferramentas escolhidas:** A utilização das ferramentas Pentaho, como o Schema Workbench, o Server e o Report Designer, demandou uma curva de aprendizado inicial. Tivemos que nos familiarizar com a modelagem dimensional e criação do esquema do Data Warehouse no Schema Workbench, bem como aprender a utilizar o Pentaho Report

Designer para a criação de relatórios personalizados e formatados adequadamente. Foi necessário dedicar tempo para estudar a documentação, buscar tutoriais e contar com o apoio da comunidade para solucionar dúvidas e superar os obstáculos encontrados.

Enfrentar essas dificuldades nos trouxe aprendizados valiosos e fortaleceu nossa capacidade de enfrentar desafios técnicos. Com perseverança e trabalho em equipe, superamos esses obstáculos e obtivemos sucesso na conclusão do projeto de Data Warehouse com base nos dados do FARS.

Referências

- Material didático disponibilizado em sala de aula.
- Tutoriais das ferramentas feitos pelos colegas de classe.