Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação

Trabalho sobre SQL Básico - Explorando a Base de Fórmula 1

Prof. Dr. Caetano Traina Júnior

Entrega dia 24 de outubro de 2018

1 Descrição das Relações

Este trabalho é feito sobre uma base que contém dados obtidos de diversas fontes na internet. Por isso, os dados não estão completamente consistentes, pois cada fonte pode codificar os dados de sua própria maneira.

O foco do trabalho é explorar dados sobre "Atentados em Escolas" (*American School Shooting*), coletados pelo jornal americano *The Washington Post* para a publicação de um artigo (https://www.washingtonpost.com/graphics/2018/local/school-shootings-database/?utm_term=.a0dc1d61a9ee) em 17 de abril de 2019. Tal coleta levantou dados desde o massacre na *Columbyne High School* em 20 de abril de 1999. Os dados utilizados para a reportagem, estão disponibilizados no *site* do jornal, e estão copiados no arquivo "`data-school-shootings-master.zip`" no repositório da disciplina. Um *script* para importar os dados para está no arquivo "`Create-SSUSA.sql`", que cria a tabela `SchoolShoots`.

O trabalho deve usar também dados sobre o censo americano mais recente, de 2010 (o próximo censo será feito em 2020). Está disponibilizado no repositório da disciplina o arquivo "`Create-SSUSA.sql`", que faz a carga de duas tabelas:

- A tabela `USCensus`, com os dados do censo americano obtidos em (https://www.census.gov/programs-surveys/decennial-census/decade.2010.html) e copiados no no arquivo "`USCities-2kCensus.txt`" no repositório da disciplina. ;
- A tabela `USStates`, com dados sobre os estados americanos copiados no no arquivo "`States.csv`".

Note que todas as tabelas contém dados corretos, mas em formatos inconsistentes. Por isso, todos os comandos em SQL devem prever essa possibilidade. Por outro lado, as tabelas contém muito mais atributos do que o necessário para responder às questões do trabalho. Assim, devem ser identificados os atributos necessários, e os comandos devem se restringir a usar/mostrar apenas os dados necessários ou que sejam importantes para o entendimento do resultado.

2 Recomendação

A primeira atividade quando se trabalha com uma base de dados é entender o que são os dados. Deve-se conhecer o conteúdo de cada tabela, localizar onde os dados estão (em quais relações, em quais atributos), e entender o formato e o significado de cada um. Para isso, é importante "ver" os dados, com uma variedade de consultas exploratórias, incluindo por exemplo:

- Ver o que cada tabela tem:
  `SELECT * FROM tabela LIMIT 10;`
- Entender o conteúdo de cada atributo que não seja de compreensão imediata.
  - Para atributos com valores, conhecer os valores:
    `SELECT Atrib, Count(*), Min(Atrib), Max(Atrib), Avg(Atrib)`

```
FROM tabela
GROUP BY Atrib -- ORDER BY Atrib;
```

  - para atributos que são identificadores de chave estrangeira, conhecer o que o significado na tabela onde ele corresponde onde o valor é chave: `SELECT TB.CE, T.C, TB.*, TB.*`

```
FROM TabelaBase TB, TabelaEstrangeira T
WHERE TB.CE, T.C;
```

- Construir tabelas temporarárias para facilitar a exploração de dados:

```
DROP TABLE IF EXISTS Temp;
SELECT TB.CE, T.C, TB.a, TB.b, TB.c, TB.x, TB.y INTO Temp
FROM TabelaBase TB, TabelaEstrangeira T
WHERE TB.CE, T.C;
```

  OBS: Criar tabelas temporárias é importante para construir modelos de análise. Se você tiver familiaridade com linguagens de análise de dados, como `R` ou `MatLab`, essas tabelas devem ser preparadas para realizar as análises (esse é o foco do *site* Kaggle).

Essas atividades não fazem parte do trabalho e não devem ser entregues, mas é importante "navegar" os dados para conhecê-los e entender como responder às tarefas propostas.

3 Definição do Trabalho

Este trabalho deve ser feito em grupo de três alunos, e entregue até o dia 29 de maio (quarta-feira). Identifique todos os participantes pelo seu nome, número USP e e-mail em um arquivo .txt.

Apenas um dos integrantes do grupo deve entregar todos os artigos do grupo. A entrega do trabalho deve ser feita pelo Tidia-4, em um arquivo em formato .zip contendo para cada tarefa:

1. O comando em SQL, em um arquivo em formato texto (.sql);
2. Um arquivo em formato pdf, contendo
   - Comentário que explique como o comando funciona (documentação do comando);
   - Identifique cada um dos atributos que são importantes para a consulta. Eles (e apenas eles) devem estar explicitados na consulta (pelo menos da cláusula *SELECT*).
   - Uma listagem que mostre o resultado da execução do comando em **Postgres**. Quando o resultado tiver até 20 tuplas, listar todas elas. Quando tiver mais de 20 tuplas, listar apenas de 10 a 20, escolhendo as tuplas mais significativas, com as tuplas de maior interesse para mostrar a corretude do comando executado. Use as cláusulas `ORDER BY` e `LIMIT` para a escolha.

Portanto, o arquivo .zip deve conter 10 arquivos, `Q01.sql` a `Q10.sql` e `Q01.pdf` a `Q10.pdf`

Os *scripts* em SQL foram criados para o SGBD PostgresQL. Caso você queira usar outro gerenciador, converta os *scripts* e os envie como arquivos adicionais na submissão dos trabalhos. Os demais arquivos de dados devem ser usados como estão.

4 Consultas em SQL

Tarefa 1) e Tarefa 2)

Estas tarefas trabalham apenas com os dados que geraram o artigo publicado pelo *The Washington Post*. Uma cópia dele está em no arquivo `Washington Post News-Shool Shootings.pdf`. Cada uma destas tarefas deve gerar uma tabela contendo todos os dados necessários para reproduzir os gráficos indicados no arquivo como "Gráfico 1"e "Gráfico 2". Os dados devem estar prontos para serem diretamente visualizados.

Tarefa 3) Verifique se as tabelas `USStates` e `USCensus` têm os dados consistentes para identificar os dados do censo das capitais de cada estado.

- Mostre um comando em SQL que mostre os dados disponíveis do censo para todas as capitais de estado.
- Modifique esse comando para que ele mostre quais dados estão inconsistentes e sugira como as tabelas devem ser corrigidas para evitar tais inconsistências.
- identifique as situações que não podem ser corrigidas.

Tarefa 4) A tabela `SchoolShoots` tem os estados identificados pelo seu nome, mas a tabela `USCensus` tem os estados identificados pela sua sigla. A tabela `USStates` tem a sigla e o nome dos estados.

- Mostre um comando em SQL que permita verificar se todos os dados estão consistentes para acrescentar a sigla dos estados na `SchoolShoots`.
  Cuidado com espaços em branco nos atributos.
- Escreva um comando simples que corrija os dados
- Modifique a tabela `SchoolShoots` para acrescentar a sigla dos estados.

$\Longrightarrow$ **Os exercícios a seguir são feitos com as tabelas corrigidas na tarefa 4.**

Tarefa 5) Baseado no que você entendeu do conteúdo das tabelas, avalie as restrições de integridade que devem existir para as três tabelas.

- Indique todas as restrições que você considera que devem existir.
- Mostre um comando SQL que verifique se cada uma dessas restrições de integridade são atendidas.
- Crie as restrições de integridade que de fato são atendidas.
- Indique o que deve ser corrigido para que as demais restrições passem a ser atendidas.

Tarefa 6) As tabela `SchoolShoots` e `USCensus` registram a distribuição de raças nas escolas, distritos onde ocorreram atentados e nas cidades correspondentes. No entanto, as raças registradas não são todas as mesmas.

- Identifique e mostre um conjunto de raças comuns a todas as distribuições.
- Escreva um comando que mostre a distribuição de raças comuns para todos os atentados, correspondentes à escola, distrito e cidade, mas normalizado pelo total de indivíduos na distribuição (divide cada valor de raça pelo total de valores naquela distribuição).

Tarefa 7) Modifique o comando anterior para acrescentar 3 atributos, que comparam cada uma das três distribuições pela soma das diferenças entre cada distribuição (distância *Manhattan*): escola×distrito, escola×cidade e distrito×cidade, ordenando decrescente por esses 3 novos atributos nessa ordem.

Tarefa 8) Mostre a população de cada raça (comum) que residem em comunidades que estão a até 10km, 50km e 200km da escola onde ocorreram atentados.

Como as coordenadas em todas as tabelas estão representadas em graus e frações decimais use a função de distância "Great Circle Distance". O arquivo arquivo `Função GCD.txt` tem essa função definida em SQL para tratar dados de tipo `NUMERIC`, e que retorna a distância em metros.

[Ultima atualização: 5 de maio de 2019]