Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
MBA em Inteligência Artificial e Big Data
– Curso 3: Administração de Dados Complexos em Larga Escala –

Questões da Quinzena 2

Prof. Dr. Caetano Traina Júnior

1 Compressão de dados

Os exercícios seguintes são baseados nas tabelas do conjunto de dados disponibilizados pelo *Repositório COVID-19 DataSharing/BR*.

Exercício 1)

Escreva um comando em SQL que crie um histograma equi-largura de distribuição das idades dos pacientes, de maneira que a largura de cada **bin** do histograma seja de '**duas idades**'.

Atente para que todas as **'idades possíveis'**, desde 0 até a maior idade registrada nos dados esteja representada no histograma.

Exercício 2)

Modifique esse comando para gerar um histograma equi-largura com 10 *bins*.

Exercício 3)

Escreva um comando em SQL que crie um histograma equi-altura com 10 *bins* da distribuição por idade, corrigindo a atribuição dos *bins* para que o histograma inclua todos os pacientes com a mesma idade no mesmo *bin*.

Exercício 4)

Escreva um comando em SQL que acrescente três novos atributos: `Numerico`, `Origem` e `Exame` na tabela `ExamLabs` que discretizem respectivamente os atributos:

1. CD_ValorReferencia, indicando se o atributo tem um valor Numérico (Numero) (possivelmente com outros textos) ou é puramente textual (Texto),

2. DE_Origem, contabilizando as origens em (Hosp)ital, (Lab)oratório, (inter)nação ou (pronto) socorro para caracterizar: (Hosp)ital, (Lab)oratório, (Atend)imento e os demais como (Outros) – use as letras entre parênteses para buscar esse padrão no atributo,

3. DE_Exame, contabilizando os exames como sendo de (Hemogr)ama, (colest)erol, e (covid) ou (pcr) ou (igm) ou (igg) para caracterizar (Hemograma), (Colesterol), (Covid) e (outros).

Use o texto entre parênteses tanto para buscar esse padrão no atributo quanto para atribuir como valor da discretização, conforme o caso.

Exercício 5)

Escreva um comando em SQL que crie o histograma tri-dimensional equi-largura de distribuição de exames (da tabela ExamLabs), tendo por dimensões:

1. DE_Hospital,

2. DE_Origem, contabilizando as origens em (Hosp)ital, (Lab)oratório, (inter)nação ou (pronto) socorro para caracterizar: (Hosp)ital, (Lab)oratório, (Atend)imento e os demais como (Outros)

3. DE_Exame, contabilizando os exames como sendo de (Hemogr)ama, (colest)erol, e (covid) ou (pcr) ou (igm) ou (igg) para caracterizar (Hemograma), (Colesterol), (Covid) e (outros).

Exercício 6)

Escreva um comando em SQL para gerar uma amostragem de aproximadamente 1% das tuplas de tal maneira que a quantidade de tuplas seja (aproximadamente) a mesma para todas as classes (<Exame, Hospital> distintos) contabilizadas no exercício anterior (use um histograma bidimensional, sem considerar a `Origem`).

Exercício 7)

Considere que a relação

`Hemograma={ID_Paciente, Basofilos, Bastonetes, Blastos, CHCM, Eosinofilos, ... VolPlaq}`

foi criada na base de dados, tal como definida no **Exercício 8** da **Primeira Lista de Exercícios** da matéria. Para cada atributo, obtenha:

- seu tipo de dado,
- a quantidade de nulos,
- a cardinalidade,
- o valor mínimo,
- o valor máximo,
- o valor médio,
- a variância,
- e o desvio padrão

2 Indexação

Exercício 8) Considere as restrições de integridade fundamentais do modelo relacional (representadas pelas restrições `PRIMARY KEY`, `UNIQUE` e `FOREIGN KEY` nos SGBDs).

1. Quais restrições usam índices para avaliar a corretude dos comandos de atualização? Porque?
2. Existe uma situação comum de demora acentuada nos comandos `DELETE` que se deve à não criação de um índice para uma das restrições fundamentais.
   - Identifique e explique qual é a situação que leva à essa demora;
   - Qual índice poderia ser criado para evitar essa demora;
   - Dê um exemplo usando a base *Repositório COVID-19 DataSharing/BR*;
   - Porque tal índice não é automaticamente criado quando a restrição é declarada.

Exercício 9) Seja o comando em SQL que cria a tabela de Exames de Hemograma, como definida no Exercício 8 da primeira lista de exercícios.

1. Crie os índices necessários para agilizar a execução desse comando.

Exercício 10) Considere um comando para associar a cada exame, a identificação do paciente, com sua idade e com o desfecho do atendimento onde o exame foi executado.

1. Escreva o comando.
2. Crie os índices necessários para agilizar a execução desse comando.
3. Existe algum índice, associado às restrições de integridade indicada pelos meta-dados da definição do *Repositório COVID-19 DataSharing/BR* que já auxiliam essa consulta? Quais e para que parte do comando?