

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO, UFMG
DCC011: INTRODUÇÃO A BANCO DE DADOS - TRABALHO PRÁTICO 2

INTRODUÇÃO

O objetivo deste trabalho é projetar e implementar um banco de dados relacional para análise de dados abertos governamentais. O projeto do banco de dados deverá seguir o processo *bottom-up* apresentado em aula. Em particular, deverão ser apresentados o diagrama entidade-relacionamento, bem como o esquema relacional referente ao banco de dados projetado. O banco de dados modelado deverá ser instanciado com dados reais. Consultas analíticas deverão ser formuladas e testadas sobre o banco de dados e discutidas quanto à sua eficiência.

Características básicas do banco de dados (até 30%): Para definição do tema, cada grupo deverá escolher um conjunto de dados publicado no ano corrente no portal <http://dados.gov.br>. Para o conjunto de dados escolhido, deverão ser apresentados o diagrama ER e o esquema relacional correspondentes, atendendo aos seguintes requisitos:

- Pelo menos 4 tipos de entidade, cada tipo com ao menos 2 atributos (além de um atributo identificador);
- Pelo menos 3 tipos de relacionamento, ao menos 1 com cardinalidade M:N;
- Pode ser necessário alterar os dados originais para atender a esses requisitos.

Consultas (até 30%): Deverão ser especificadas e executadas um total de 10 consultas em SQL, sendo:

- 2 consultas envolvendo as operações de seleção e projeção;
- 3 consultas envolvendo a junção de duas relações;
- 3 consultas envolvendo a junção de três ou mais relações;
- 2 consultas envolvendo funções de agregação sobre o resultado da junção de pelo menos duas relações.

Características avançadas (até 40%): Para avaliação de eficiência, cada uma das 10 consultas especificadas deverá ser formulada de 2 formas diferentes considerando, por exemplo, opções como utilização ou não do operador JOIN, utilização ou não de consultas aninhadas, criação de índices para determinadas colunas, entre outras.

OBSERVAÇÕES:

1. O trabalho deverá ser feito em **grupo de 3-4 pessoas**.
2. O SGBD **SQLite** embarcado em um **Jupyter notebook** deverá ser utilizado para implementação do banco de dados e execução das consultas. **DICAS:** Para acelerar a inserção de dados, desative temporariamente a validação de chaves estrangeiras (**PRAGMA foreign_keys = 0**). Para avaliação de tempo de resposta, desative o mecanismo de *cache* (**PRAGMA cache_size = 0**) a fim de evitar efeitos quanto à ordem de execução das consultas. Para minimizar outros possíveis efeitos, reporte o tempo médio sobre múltiplas execuções de cada consulta (e.g., preceda cada consulta com o comando mágico **%timeit**, ou cada célula com o comando **%%timeit**).
3. A avaliação compreenderá o relatório final entregue e a apresentação em aula. Além da funcionalidade básica prevista, a avaliação irá considerar a criatividade e a eficiência das consultas formuladas.

CALENDÁRIO:

16/10 (qua): Entrega da proposta: definição do grupo e do tema (**PDF, máx. 1 página**)

A proposta deverá descrever o conjunto de dados escolhido do portal <http://dados.gov.br/>. Para evitar temas duplicados, informe seu grupo e dataset escolhido na seguinte planilha: <http://bit.ly/2ONulxw>. Na proposta escrita (1 página), deverão ser descritas as entidades e relacionamentos presentes, o número de instâncias de cada entidade e relacionamento, e possíveis consultas a serem formuladas sobre esses dados.

04/11 (seg): Entrega do relatório final (**verifique o template no Moodle**)

O relatório final deve ser feito no formato notebook (.ipynb), incluindo as seguintes seções (um template está disponível na atividade criada no Moodle): (1) Título; (2) Membros (nome e matrícula); (3) Descrição dos dados (como foram processados?); (4) Diagrama ER; (5) Diagrama relacional; (6) Consultas; (7) Auto-avaliação dos membros

Idealmente, o processamento dos dados deverá ser feito no próprio notebook e implementado na Seção 3.

Os diagramas ER e relacional produzidos deverão ser anexados nas Seções 4 e 5, respectivamente.

Na Seção 6, cada consulta deverá ser definida através de uma explicação textual e do respectivo comando SQL, acompanhado do resultado da consulta e do tempo de execução em comparação a consultas alternativas. Gráficos ilustrativos também podem ser incluídos para facilitar a visualização dos resultados.

Na Seção 7, a auto-avaliação deverá descrever as atividades realizadas individualmente por cada membro.

04,06,11,13/11: Apresentação dos trabalhos

Cada grupo terá em torno de 15 min (a confirmar, dependendo do número de grupos) para apresentar o tema do trabalho, a modelagem desenvolvida (conceitual e relacional), uma seleção das várias consultas formuladas e de seus resultados, e uma descrição do trabalho desenvolvido por cada membro do grupo. Para apresentação, o notebook desenvolvido e entregue deverá ser utilizado para demonstração ao vivo.