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 um processo *bottom-up*, iniciado a partir da análise de um conjunto de dados existente, passando pela normalização de seu esquema relacional e mapeamento para o esquema ER correspondente. O conjunto de dados reais escolhido deverá ser então inserido no banco de dados normalizado. Por fim, consultas analíticas deverão ser formuladas e testadas sobre esse 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 **criado a partir de 2021** no portal dados.gov.br/dataset. Para o conjunto de dados escolhido, deverão ser apresentados o **esquema relacional normalizado** e o **diagrama ER** 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** (perfazendo 20 consultas ao todo) considerando, por exemplo, utilização ou não do operador JOIN ou de consultas aninhadas, criação de índices para determinadas colunas, entre outras.

OBSERVAÇÕES:

1. O trabalho deverá ser feito em **grupo de 3-4 alunos.**
2. O SGBD **SQLite** embarcado em um **notebook** no Google Colab 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 *caching* **(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 (notebook) final entregue e a apresentação gravada. Além da funcionalidade básica prevista, a avaliação irá considerar a criatividade e a eficiência das consultas formuladas.

CALENDÁRIO:

26/01 (qua): Entrega da proposta (.pdf, máx. 1 página, via Moodle): definição do grupo (3-4 alunos) e do tema

A proposta deverá descrever o conjunto de dados escolhido do portal dados.gov.br/dataset. Para evitar temas duplicados, informe seu grupo e dataset escolhido na seguinte planilha: https://bit.ly/33DAmfw. Na proposta escrita (1 página), deverão ser listadas as entidades e relacionamentos presentes, o número de instâncias de cada entidade e relacionamento, e possíveis consultas a serem formuladas sobre esses dados. A proposta deve ser submetida via Moodle.

16/02 (qua): Entrega do relatório final (.ipynb, 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) Autoavaliação 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 autoavaliação deverá descrever as atividades realizadas individualmente por cada membro.

16/02 (qua): Submissão das apresentações (vídeos submetidos via YouTube, links via Moodle)

*Cada grupo deverá preparar uma apresentação sucinta do trabalho realizado, incluindo uma descrição do tema, da modelagem desenvolvida (esquemas ER e relacional), e uma seleção das várias consultas formuladas (incluindo sua explicação em alto nível e seus resultados retornados). A apresentação deverá ser gravada em vídeo (**máx. 6 min**; dica: utilize o software https://www.loom.com/). Um link (unlisted) para a apresentação no YouTube deverá ser submetido para avaliação via Moodle. Todos os membros do grupo deverão participar do vídeo, identificando-se e apresentando as partes que lhes couberam.*