Trabalho Final – Etapa 6 – Parcial 2

MC536 – Bancos de Dados: Teoria e Prática Instituto de Computação Universidade Estadual de Campinas		2º semestre de 2021 Turma A Professor: André Santanchè

Resumo

O objetivo geral do projeto final desta disciplina é tornar disponível publicamente através do Github um Dataset cujo conteúdo possa ser usado para análise de dados e pesquisa científica.

Descrição Geral

Nesta etapa, será apresentada a primeira versão da base, incluindo a especificação das fontes de dados, modelos e processos de transformação, tratamento e integração.

Devem ser apresentadas as primeiras versões dos seguintes elementos:

- Modelo conceitual e respectivos modelos lógicos (considerando que deve haver mais de um) do dataset publicado.
- Operações realizadas para a construção do dataset, dentre as seguintes:
  - extração de dados de fontes não estruturadas como, por exemplo, páginas Web
  - agregação de dados fragmentados obtidos a partir de API
  - integração de dados de múltiplas fontes
  - tratamento de dados
    - por exemplo, tratamento de dados faltantes, incompletos e inconsistentes
  - transformação de dados para facilitar análise e pesquisa
- Listagem de perguntas de análise/pesquisa a ser aplicadas ao dataset.
- A partir das perguntas feitas, exemplos de consultas que as respondem. Para o conjunto de perguntas, devem ser considerados, no mínimo, dois modelos lógicos (e.g., relacional e documentos) e idealmente três modelos lógicos (e.g., relacional, documentos e redes). Cada pergunta será respondida pelo acesso da base em um dos modelos lógicos, entretanto, a equipe pode responder a mesma pergunta de duas maneiras diferentes usando dois modelos diferentes.
  - Cada modelo lógico deve cumprir um papel dentro da análise mais adequado ao seu modelo, por exemplo, o modelo de grafos pode ser aplicado a uma rede de conhecimento, ou rede social; o modelo de documentos pode ser aplicado a partes da análise que envolvam documentos semiestruturados; o modelo tabular a dados mais estruturados.

Deve ser detalhada a primeira versão dos modelos (conceitual e lógico) que a equipe usará. Trata-se um modelo preliminar, que poderá ser modificado no futuro, serve para a equipe mostrar o ponto de partida para as análises pretendidas.

Também deve ser escolhido um dos modelos lógicos para a entrega de uma primeira versão de perguntas/consultas. A equipe pode escolher qualquer um dos modelos (tabelas, hierárquico ou rede), mas sugere-se o modelo relacional (tabelas), pois foi o tratado com mais detalhes em sala. Devem ser observados aspectos de normalização no modelo lógico.

Para o modelo escolhido, deve ser apresentado um primeiro programa que realiza as operações de preparo do dataset (extração, agregação, integração, tratamento, transformação).

As consultas que respondem às perguntas devem usar algum tipo de query, ou seja, não

deve ser feito processamento completamente em memória. Se for escolhido modelo tabular, a sugestão é o uso de SQL.

Por ser um primeiro conjunto de consultas, ele poderá ser refinado e expandido posteriormente. Uma sugestão é um conjunto de, pelo menos, cinco queries de complexidade média em SQL que apresente resultados interessantes de uso do dataset.

Entrega e Apresentação

Deve ser postado no projeto do Github da equipe.

Esta entrega é composta de dois momentos:

(1) disponibilização da descrição da proposta e respectivos slides em projeto do projeto no Github até 5 de novembro até 7h59; conforme template.

(2) apresentação e arguição da proposta dia 5 de novembro dentro do horário da aula e laboratório segundo ordem de arguição a ser divulgada.

A apresentação da proposta deve ter duração máxima de 8 minutos. Todos os membros da equipe devem participar da apresentação e arguição subsequente.

A proposta (texto e slides) deve contemplar, pelo menos, os seguintes aspectos:

- descrição do tema do dataset, incluindo motivação e contexto gerador;
- modelo conceitual do dataset;
- modelos lógicos do dataset;
- fontes de dados e operações de preparo do dataset (extração, agregação, integração, tratamento, transformação);
- perguntas de pesquisa/análise que podem ser respondidas pelo dataset;
- primeiro conjunto de consultas.