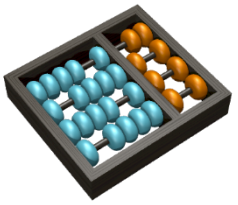
Universidade Estadual de Campinas
Instituto de Computação



Cursos: Bacharelado em Ciência e Engenharia da Computação

Trabalho de MC536 - Banco de Dados

Nomes: *Pedro Barros Bastos RA: 204481*
Rafael Cabral Pili RA: 185999
Gabriel Volpato Giliotti RA: 197569

Campinas – SP
2019

Descrição Geral do Projeto

O projeto descrito e desenvolvido no decorrer do semestre apresentou quatro fases de desenvolvimento, sendo elas:

- Fase 1: Pesquisa de bases de dados direcionadas à área da saúde na rede, para realização de análises. (Os alunos podiam trocar para outras bases durante as fases seguintes).

- Fase 2a: Escolha das duplas e da base de dados para prosseguir para a fase seguinte. Votação realizada para escolher as melhores bases de dados da Fase 1.

- Fase 2b: Fase de engenharia reversa, onde foram construídos modelos Entidade-Relacionamento e Orientado à Objetos com conceitos aprendidos em aula.

- Fase 3: Construção e apresentação de consultas em SQL em uma base escolhida na Web (de preferência uma das escolhidas na Fase 1), sobre um tema mais específico da área da saúde , junto ao Modelo Relacional da base.

- Fase 4: Construção e apresentação de consultas em XQuery sobre uma Base de dados hierárquica, focando nas vantagens e desvantagens entre os tipos de estrutura estudados até então

- Fase Final: Construção e apresentação de consultas em SQL, XQuery e Cypher (análise de redes), bem como uma apresentação final do trabalho em slides e revisão das fases anteriores buscando melhor entendimento dos diferentes tipos de estrutura.

Fases 1 e 2

Durante a fase 1 do projeto, foi apresentada uma introdução sobre bancos de dados, assim como diferentes tipos de bancos de dados existentes na área da saúde que podem ser explorados e foram feitas algumas atividades de avaliação desses bancos. As atividades foram passadas como tarefas em aula ou laboratório, para pesquisa e armazenamento de referências para futura seleção . Em seguida, já na fase 2, formamos duplas com nomes de grupos (CPP e OPD) e escolhemos as bases por um sistema de votação, onde as escolhas das bases deveriam convergir para alguns bancos melhores estruturados.

Em seguida, ainda em fase de duplas, foi realizada uma atividade de engenharia reversa onde tomamos uma base de dados sobre saúde e, com conceitos aprendidos em aula sobre o Modelo Entidade-Relacionamento, criamos diagramas ER e Orientado à Objetos (Entidade Relacionamento e UML).

Fase 3 e 4

Nesse ponto, aplicando conceitos da linguagem SQL aprendidos em aula junto do Modelo Entidade-Relacionamento, fizemos consultas de análise Exploratória sobre os dados do base Protein Atlas (que escolhemos e apresentamos os modelos ER e OO na fase anterior), buscando encontrar relações entre anticorpos, tecidos, amostras fornecidas de rna, tipagem sanguínea e gênero dos pacientes, onde poderia haver algum ponto interessante a ser notado em meio a grande quantidade de dados apresentados. Com esse cenário, ainda fizemos uma apresentação da Fase 3 do projeto, de 5 minutos em sala, para receber as considerações do professor quanto ao andamento melhorias sobre o projeto.

Logo em seguida, durante as aulas e os laboratórios, foram apresentados novos meios de estruturação dos dados que poderiam ser utilizados. Para isso, diferentes tipos de organizações de arquivos, com suas vantagens e desvantagens, foram exibidos e posteriormente tiveram atividades para suas aplicações em laboratório. Em especial, tivemos a estrutura Hierárquica (ou de hierarquia) de dados apresentada, onde através da linguagem de XQuery chamada XPath, podemos realizar consultas em diferentes níveis da estrutura. Finalmente, na Fase 4 do projeto, houve a união de duplas entre os alunos da sala, formando novos grupos, e assim, realizamos uma segunda apresentação de 5 minutos, em sala, agora com um olhar comparativo entre XQuery e SQL e quais a vantagens e desvantagens de se aplicar esses diferentes tipos de estruturas (Hierarquias e tabelas) nas análises de dados.

Fase Final

Finalmente, na Fase Final, tomamos um novo tipo de estrutura dado em laboratório chamado Cypher, que é uma estruturação dos dados em forma de grafos para realização de diferentes tipos de análises de redes. Junto a uma revisão e reestruturação dos outros tipos de linguagens e estruturas, construímos um repositório no GitHub com a apresentação final do projeto, onde constam os dados analisados, e cada tipo de estrutura recebe 5 queries (5 de SQL, 5 de XPath e 5 de Cypher), além desse relatório e a união de todas a outras fases do projeto de todos os integrantes do grupo, mais uma apresentação de até 30 minutos em sala.

O objetivo aqui, além de apresentar a fase final do projeto, é entender que os diferentes tipos de estruturas (Tabelas, hierarquias ou grafos) que podemos empregar na análise de grandes quantidades de dados. Sendo assim, diferentes perspectivas sobre os dados surgem e devemos escolher a melhor forma de análise, tanto para otimização como para facilidade de acesso, visualização entre outros pontos. Podemos citar alguns exemplos como a análise de redes ou comunidades que podem ser melhor realizadas com dados estruturados em forma de grafo, ou ainda buscas mais eficientes em estruturas hierárquicas, que facilitam o acesso com consultas relativamente simples para diferentes níveis da hierarquia.

Portanto, esse projeto tem por objetivo a exposição de pontos importantes que vão além da sala de aula, buscando mostrar os mais novos e diferentes modos de se estruturar um conjunto grande de dados, além de oferecer um comparativo para os diferentes tipos de estruturas e realizar análises confiáveis e bem definidas sobre dados do cotidiano.