MAC0426 / MAC5760

Projeto: Avaliação de Desempenho de Bancos de Dados

Prazo para a definição do grupo e do BD: 07/10/2022
Prazo para a entrega do projeto: 04/11/2022

1 Descrição Geral do Projeto

Neste projeto, a tarefa será fazer um estudo para medir e comparar o desempenho dos SGBD PostgreSQL e o do [MongoDB **ou** Neo4j] na execução de diversos tipos de operações sobre um grande banco de dados, com e sem o apoio de índices. Ou seja, o grupo avaliará duas implementações diferentes de um mesmo BD: uma no modelo relacional (no PostgreSQL) e outra em um não-relacional, escolhendo entre o modelo de documentos (no MongoDB) ou o modelo de grafos (no Neo4J). No final do projeto, o grupo deverá entregar:

- um relatório descrevendo o estudo realizado e os seus resultados;
- um link para um vídeo de 10 minutos apresentando o trabalho;
- um link para um repositório online com os códigos-fonte para a criação dos BDs, execução das operações, arquivos com os resultados das medições de desempenho realizadas, geração dos gráficos, etc.

O grupo deverá escolher (ou criar) o BD a ser utilizado no estudo. Pode-se utilizar *datasets* disponíveis publicamente na Internet como base para a criação do BD, contanto que isso seja explicitamente indicado no trabalho. A referência à fonte dos dados deve ser incluída no relatório.

2 Experimentos de Avaliação de Desempenho

Para os experimentos de avaliação de desempenho, primeiro, para cada um dos dois SGBDs, o grupo terá que projetar um esquema para o BD no modelo de dados implementado pelo SGBD, criar o BD e povoá-lo com os dados baixados. Depois, o grupo deverá especificar um conjunto de operações de modificação e recuperação de dados sobre o BD e escrevê-las com comandos nas linguagens dos dois SGBDs. Após a definição das operações, o grupo deverá identificar índices que possam beneficiar as operações de consulta.

As operações a serem avaliadas deverão incluir:

- Inserções de dados
- Remoções de dados

- Alterações de dados
- Buscas de registros pelo seu(s) atributo(s) chave-primária
- Buscas de registros por diferentes tipos de atributos não-chave, inclusive com condições de seleção compostas
- Buscas envolvendo relacionamentos (junções e junções de junções)
- Buscas envolvendo operações de agrupamento e agregações (ex.: soma, média, mínimo, máximo, contagem)

O grupo precisará projetar experimentos para avaliar o desempenho das operações em diferentes cenários. Um cenário que deve existir é o do BD sem índices. Outros cenários podem ser derivados a partir da criação de índices sobre diferentes atributos e índices de diferentes tipos (como Árvore B+ e Hash). Cada cenário terá os seus resultados de desempenho e é a comparação desses resultados que deve ser analisada no estudo.

Cada operação definida precisará ser executada diversas vezes em cada cenário, pois o tempo de execução dela pode variar de uma execução para outra (devido a motivos diversos, tais como carga na máquina no momento da execução, paginação do SO e cache do SGBD). Nos seus experimentos, tente evitar tanto quanto possível que fatores externos interfiram nos resultados dos experimentos. Tente isolar o ambiente de execução, desabilitar caches e usar clientes de conexão "leves" para submeter as operações aos SGBDs. Para cada SGBD e cada cenário, execute cada operação pelo menos 20 vezes. Todos os experimentos deverão ser executados numa mesma máquina, caso contrário, seus resultados não serão comparáveis.

3 Relatório com os Resultados do Projeto

O relatório deve descrever o BD escolhido, o seu esquema relacional e o não-relacional, as consultas definidas, os experimentos e cenários projetados, o ambiente computacional usado na execução (que inclui configurações da máquina e dos softwares dos SGDBs), as medidas de desempenho coletadas das execuções e as conclusões a que se chega a partir dos resultados.

É obrigatório na escrita do relatório o uso de diagramas, gráficos e tabelas feitas pelo grupo para apresentar os esquemas e mostrar e comparar o desempenho das operações. Para cada operação e SGBD, apresente o tempo médio de execução da operação e medidas variabilidade/dispersão (ex.: desvio padrão ou intervalo de confiança). Adicionalmente, recomenda-se apresentar todos os tempos das execuções de uma operação como um gráfico de linha (com o tempo no eixo y e o número da execução no eixo x) ou um diagrama de caixa (*boxplot*), possibilitando a visualização completa dos resultados obtidos.

A métrica de desempenho que deve ser obrigatoriamente avaliada nos experimentos é o tempo de execução das operações. Mas outras métricas de desempenho também podem ser consideradas pelo grupo, tais como uso de espaço em disco, o uso de memória RAM e ocupação da CPU.

O seguinte modelo sugere uma organização para o relatório, mas ela não é obrigatória. O importante é que o relatório contenha todos os itens listados acima:

`https://docs.google.com/document/d/1O3J_NG71ZcUTQv5MeUEvc3gpd8K1xISBuikDWG58_tw/edit?usp=sharing`

Inclua citações a referências bibliográficas no texto, para indicar as fontes usadas como base para o estudo realizado. É importante destacar que as referências devem ser utilizadas apenas para auxiliar o

desenvolvimento do estudo e a escrita. Ou seja, **o trabalho precisa ser desenvolvido inteiramente pelo grupo**. Conteúdo copiado sem a devida referência será considerado plágio.

4 Entrega

O relatório e os links para o vídeo de apresentação e para o repositório do projeto deverão ser entregues até às 23:59 do dia 07/11/2022.

Avisos importantes:

- No README do repositório de código do projeto, deve haver instruções para a execução dos scripts/programas criados.
- Os códigos e as instruções disponibilizados devem possibilitar a reprodução do estudo realizado. Na avaliação da entrega, a professora irá reexecutar os experimentos.
- Não deixe o repositório de código do projeto aberto para acesso de qualquer pessoa. Infelizmente, tem sido comum ocorrências de alunos que plagiam trabalhos obtidos de repositórios abertos. Na entrega do projeto, conceda acesso à professora ao repositório fechado do grupo. Tanto na plataforma github quanto na gitlab, a conta da professora está associada ao e-mail `kellyrb@ime.usp.br`.