EACH | campus capital USP LESTE
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Programa de Pós-graduação em Sistemas de Informação
PPgSI

VI Workshop de Dissertações de Mestrado do PPgSI 2019

Indexação de Dados em Bases com Elevado Volume de Dados

Autoria de:	Mariângela Ferreira Fuentes Molina			
Orientação de:	Prof. Dr. Marcelo Medeiros Eler			
Coorientação de:				
Linha de pesquisa:	[X] Gestão e Desenvolvimento de Sistemas		[ ] Inteligência de Sistemas	
Área de pesquisa:	[X] Banco de dados [ ] Gestão de tecnologia da informação	[ ] Engenharia de software [ ] Interação humano computador	[ ] Inteligência artificial [ ] Reconhecimento de padrões	[ ] Processamento gráfico [ ] Otimização
Área de aplicação:	[ ] Ambientes corporativos / Processos de negócio [ ] Economia [ ] Jogos / Jogos sérios [ ] Saúde	[ ] Bioinformática [ ] Educação / Educação a distância [ ] Linguística / Língua natural [ ] Outra Qual? _____	[ ] Biometria [ ] Governo eletrônico [ ] Quimioinformática [X] Geral	[ ] Dispositivos móveis [ ] Internet / Redes sociais [ ] Robótica
Semestre no curso (na data do workshop):	[ ] 2º semestre	[X] 3º semestre	[ ] 4º semestre	[ ] 5º semestre
Qualificação:	[ ] Qualificação já realizada em: dd/mm/aaaa		[X] Realização da qualificação planejada para: 07/10/2019	
Defesa:	Prazo máximo para depósito: 01/02/2021		Realização da defesa planejada para: 06/07/2020	
Publicações associadas ao projeto de mestrado:	Sem publicações até o momento.			

Resumo do projeto de pesquisa

Contexto:

Sistemas de bancos de dados são utilizados desde a década de 1960 com o surgimento do modelo hierárquico para armazenamento e recuperação de dados das aplicações. O modelo relacional proposto por Edgar F. Cood no artigo "Um Modelo Relacional de Dados para Grandes Bancos de Dados Compartilhados", predomina no mercado atual de sistemas gerenciadores de bancos de dados, sendo o mais utilizado por empresas em todo o mundo, fornecendo características importantes tais como consistência dos dados e acesso simultâneo aos mesmos por múltiplos usuários. Mesmo com a consagração desse modelo, os modelos de dados continuaram a evoluir, passando pelo modelo orientado a objetos e atualmente com os modelos denominados NoSQL, cuja estrutura de armazenamento fogem aos conceitos relacionais.

Com o passar dos anos, entretanto, o volume dos dados gerados e armazenados tem sofrido um crescimento muito grande, chegando na casa de 2,5 Zetabytes diáiros de dados. Acredita-se que este fenômeno tenha relação direta com a grande utilização de redes sociais, compras na internet e pesquisas em sites de busca, que normalmente são fontes para processos de análise.

Os sistemas de bancos de dados atuais, porém, não se adaptaram totalmente a essa nova realidade, e nem sempre conseguem operar com um desempenho adequado para as aplicações que deles dependem. Como consequência, as consultas ficaram seriamente comprometidas com relação ao tempo de resposta para os métodos de armazenamento de dados existentes, sendo tema para diversos trabalhos de pesquisa.

Ao analisar bases com elevado volume de dados é possível observar a existência de um padrão nas consultas realizadas com maior frequência, uma vez que informações para tomada de decisão de usuários são especificas de cada domínio de negócio. Para estas consultas damos o nome de consultas específicas (specifc queries). Essas consultas por sua vez são capazes de consumir grandes recursos de hardware por possuírem uma complexidade alta em seus algoritmos, fazendo-se necessário mais investimentos para aquisição de componentes que agilizem seu processamento.

Problema de pesquisa:

Nas técnicas tradicionais de modelagem dos bancos de dados, analistas e programadores de aplicação tendem a modelar o banco de dados de modo que suas operações sejam processadas em um tempo mínimo e com um consumo mínimo de recursos, não observando a diferença na prioridade de algumas consultas em detrimento de outras.

Dentre as novas abordagens encontradas durante a revisão da literatura, as quais visam aumentar a eficiência dos bancos de dados com volume elevado, não foi encontrada uma proposta cuja modelagem privilegie os dados que são acessados a partir de consultas específicas predefinidas, as quais são mais comuns e possuem importância maior para o sistema.

Objetivo de pesquisa:

O objetivo geral desta pesquisa é desenvolver uma nova abordagem para indexação de dados onde consultas específicas possam ser respondidas com alto desempenho em bases com elevado volume de dados. A abordagem proposta endereça uma solução para que custos com infraestrutura de armazenamento de dados possam ser reduzidos, uma vez que se almeja que a abordagem proposta possibilite a indexação dos dados sob demanda, possibilitando que recursos de hardware sejam utilizados conforme necessário em uma estrutura de computação em nuvem.

Caracterização da solução em desenvolvimento:

A solução que está sendo elaborada visa fazer inicialmente a identificação das consultas específicas em duas bases de dados experimentais e, a partir dessa identificação, pretende-se criar duas estruturas para indexação dos dados, sendo a primeira voltada para os dados co-acessados nas consultas específicas e a segunda para os demais dados, utilizando técnicas de clusterização das estruturas de índices para redução do espaço de busca, que reduzirá o tempo de acesso aos dados e a quantidade de processamento envolvido, quando comparado ao modelo de dados tradicional. A estrutura dos índices está sendo estudada, assim como algoritmos para percorrer estruturas de dados em árvores e grafos, para que o modelo proposto possa ser implementado de forma que as buscas ocorram eficientemente, fazendo com que o dado que esteja sendo buscado seja encontrado com tempo pequeno a um custo computacional baixo.

As bases de dados serão submetidas a testes de desempenho, onde o tempo de processamento e o consumo de recursos de hardware serão medidos, comparando-se a execução de um conjunto de consultas específicas na solução proposta e em um sistema de banco de dados tradicional para validação dos resultados.

Fundamentos:

Clusterização de arquivos de índice: consiste na segmentação dos arquivos que contém as páginas de indices para busca dos dados no banco de dados

Algorítmos de ordenação e busca: são algoritmos consagrados que podem fazer com que tarefas não triviais possam ser executadas em um tempo menor ou com menor gasto computacional..

Trabalhos relacionados:

- ALUC , G.; • OZSU, M. T.; DAUDJEE, K. Building self-clustering rdf databases using tunable-lsh. The VLDB Journal – The International Journal on Very Large Data Bases, Springer-Verlag, v. 28, n. 2, p. 173 - 195, 2019.

- KUMAR, B. S.; REDDY, H. V.; RAJU, T. A.; VENNAM, P. Clustering categorical data using rough membership function. In: IEEE. 2014 International Conference on Computational Intelligence and Communication Networks. [S.l.], 2014. p. 602 - 607.
- RAHMAN, M. H.; ABID, F. B. A.; ZAMAN, M.; AKHTAR, M. N. Optimizing and enhancing performance of database engine using data clustering technique. In: IEEE. 2015 International Conference on Advances in Electrical Engineering (ICAEE). [S.l.], 2015. p. 198 - 201.

Validação:
A avaliação dos resultados da pesquisa consiste na verificação da eficiência do modelo proposto em comparação à baseline principal. Utilizando-se duas bases de dados de domínio público ou bases reais descaracterizadas, será realizada uma série de comparações, onde serão verificados o tempo de resposta, o consumo de memória e processamento para um conjunto de consultas. As comparações serão feitas utilizando-se uma modelagem na solução proposta no trabalho de pesquisa e outra modelagem seguindo o padrão da baseline atual.

Limitações, riscos e ameaças:
A presente pesquisa tem como escopo o desenvolvimento de uma abordagem para indexação de dados co-acessados por consultas específicas, restringindo-se a melhorar o desempenho exclusivamente para essas consultas. Outras consultas que não fazem parte do conjunto específico escolhido não serão testadas ou farão parte dos resultados finais.
O tempo pode ser uma ameaça à execução de todas as consultas desejadas, dependendo de sua complexidade e do tempo que cada solução levará para computar os resultados.

Contribuição científica:
O trabalho traz como contribuição científica uma nova abordagem implementada e testada para ser utilizada em ambientes com bases de dados com elevado volume de dados, com conjunto bem definido de consultas específicas realizadas por suas aplicações.

Contribuição tecnológica (se pertinente):
Como contribuição tecnológica, o presente trabalho traz um conjunto de testes que medem o tempo de resposta e consumo de hardware para um conjunto de consultas no modelo de dados tido como baseline

Método de pesquisa			
Gênero (escolha UM)	[ ] Pesquisa teórica	[ ] Pesquisa prática	[X] Pesquisa empírica [ ] Pesquisa metodológica
Natureza (escolha UMA)	[ ] Pesquisa básica		[X] Pesquisa aplicada
Abordagem (escolha UMA)	[X] Pesquisa quantitativa	[ ] Pesquisa qualitativa	[ ] Pesquisa mista (quali-quanti)
Revisão de literatura* (você pode escolher mais de uma)	[X] Revisão narrativa [X] Revisão descritiva [ ] Revisão de escopo	[ ] Meta-análise [ ] Revisão sistemática qualitativa [ ] Revisão *guarda-chuva*	[ ] Revisão teórica [ ] Revisão realística [ ] Revisão crítica
Procedimento técnico principal (escolha UM)	[X] Pesquisa experimental [ ] Pesquisa bibliográfica [ ] Pesquisa documental [ ] Pesquisa *ex-post-facto*	[ ] Pesquisa com *survey* [ ] Estudo de caso [ ] Pesquisa participante [ ] Pesquisa-ação	[ ] Pesquisa etnográfica [ ] Teoria fundamentada em dados [ ] Ciência do projeto [ ] Outra Qual? _____
Análise de dados (você pode escolher mais de uma)	[X] Estatística descritiva [ ] Estatística inferencial	[ ] Teste estatístico [ ] Análise de conteúdo	[ ] Análise do discurso [ ] Outros: _____

* Definição de tipos de revisões de literatura estabelecida por Paré, G., Trudel M-C., Jaana M., Kitsiou, S. Synthesizing Information systems knowledge: A typology of literature reviews. In: Information & Management 52, p. 183-199, 2015. DOI: 10.1016/j.im.2014.08.008

Próximas atividades:
Amadurecimento da abordagem de estruturas indexadas; Implementação da solução proposta;
Execução dos testes com a solução implementada e com a baseline principal; Sumarização e análise dos resultados

Opcional: Forneça um esquema gráfico que mostre aspectos de sua pesquisa. Por exemplo: um fluxograma para construção da sua solução ou um infográfico para sua proposta de pesquisa. Se necessário, use a quarta página.