UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO TECNOLÓGICO
DEPARTAMENTO DE ENGENHARIA ELÉTRICA
PROJETO DE GRADUAÇÃO

MATEUS BENTO ALVES VASCONCELLOS

ANÁLISE DE SISTEMAS GERENCIADORES DE BANCOS DE DADOS (SGBD) PARA ARMAZENAMENTO DE UMA QUANTIDADE VOLUMOSA DE GRAFOS

VITÓRIA-ES

MARÇO/2022

Mateus Bento Alves Vasconcellos

ANÁLISE DE SISTEMAS GERENCIADORES DE BANCOS DE DADOS (SGBD) PARA ARMAZENAMENTO DE UMA QUANTIDADE VOLUMOSA DE GRAFOS

Parte manuscrita do Projeto de Graduação do aluno Mateus Bento Alves Vasconcellos, apresentado ao Departamento de Engenharia Elétrica do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Engenheiro Eletricista.

Vitória-ES

Março/2022

Mateus Bento Alves Vasconcellos

ANÁLISE DE SISTEMAS GERENCIADORES DE BANCOS DE DADOS (SGBD) PARA ARMAZENAMENTO DE UMA QUANTIDADE VOLUMOSA DE GRAFOS

Parte manuscrita do Projeto de Graduação do aluno Mateus Bento Alves Vasconcellos, apresentado ao Departamento de Engenharia Elétrica do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Engenheiro Eletricista.

Profa. Dra. Marcia Helena Moreira Paiva
Universidade Federal do Espírito Santo
Professor da Disciplina

Profa. Dra. Marcia Helena Moreira Paiva
Universidade Federal do Espírito Santo
Orientador

Mateus Bento Alves Vasconcellos
Universidade Federal do Espírito Santo
Aluno

Vitória-ES

Março/2022

RESUMO

Diversas indústrias, comércios e a área acadêmica utilizam bancos de dados para armazenar as mais variadas formas de dados, gerando uma demanda cada vez maior por métodos de melhoras de desempenho de todos os tipos. Surge, portanto, o desafio de encontrar as otimizações mais eficazes e os melhores gerenciadores de bancos de dados para reduzir o tempo de consulta e volume para cada uso específico. Ao mesmo tempo, a Teoria de Grafos é cada vez mais utilizada como um modelo altamente capaz de resolver, visualizar e armazenar problemas matemáticos, da engenharia, da computação e da indústria pelas suas diversas aplicações práticas. Este trabalho busca estudar e aplicar testes de desempenho nos principais SGBD (Sistemas Gerenciadores de Bancos de Dados) realizando a comparação de desempenho de consulta e volume utilizado para os principais tipos de consultas a bancos de dados de grafos.

Palavras-chave : Base de Dados. Desempenho. Grafos. Sistemas Gerenciadores.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CPU	*Central Processing Unit*
CPID	Centro de Pesquisa, Inovação e Desenvolvimento
GPU	*Graphics Processing Unit*
RNP	Rede Nacional de Ensino e Pesquisa
SGBD	Sistemas Gerenciadores de Banco de Dados
SQL	*Structured Query Language*
NoSQL	*Not Only Structured Query Language*
UFES	Universidade Federal do Espírito Santo

SUMÁRIO

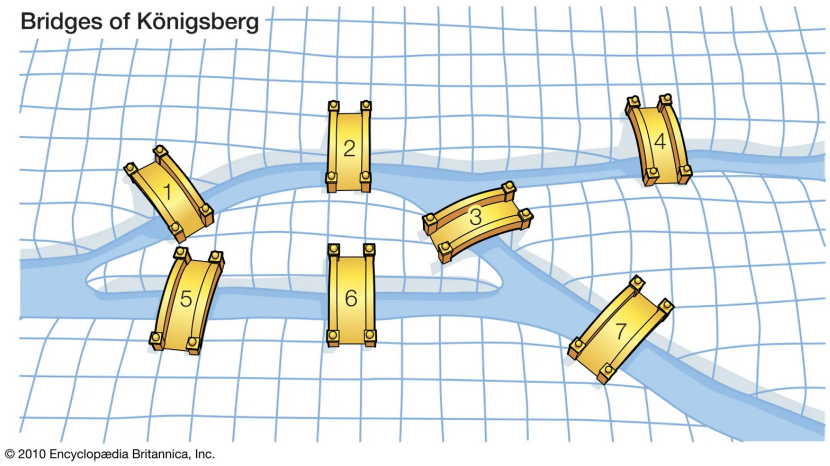
1 INTRODUÇÃO

A pesquisa e a otimização de bancos de dados são assuntos em alta na área da computação há décadas, estando presentes em todos os tipos de sistemas e aplicativos utilizados mundialmente. No entanto, conforme a indústria se solidifica e amadurece, a competitividade e a necessidade de sistemas mais eficientes também aumentam.

Um problema de bancos de dados pode ser entendido como um problema de grafos quando analisa-se não só os valores a serem armazenados, mas suas conexões e relações (GUBICHEV, 2015).

O primeiro artigo sobre grafos foi escrito pelo matemático Leonard Euler, em 1736, enquanto tentava descobrir se era possível atravessar as sete pontes da cidade de *Königsberg*, na Prússia, sem repetir nenhuma (BIGGS E. KEITH LLOYD, 1986). A Figura 1 ilustra o problema. Euler abstraiu os possíveis caminhos das pontes em retas e suas intersecções em pontos, criando, talvez, o primeiro grafo da história. Desde então, o tema tem ganhado cada vez mais relevância, por ser altamente utilizado para abstração de relações entre objetos, principalmente nos âmbitos da computação e telecomunicações.

Figura 1 – Pontes de *Königsberg*



Fonte: Carlson (2010)

O crescimento da Teoria de Grafos e a necessidade de modelar sistemas complexos e volumosos de forma eficiente torna possível a modelagem de redes de telecomunicações utilizando grafos, onde nós são representados por pontos interligados na rede e arestas as suas ligações. No entanto, pela necessidade de rapidez da análise dos dados armazenados, e, lidando com grandes volumes de dados, faz-se preciso identificar o SGBD mais veloz para consulta das principais características no contexto de redes de telecomunicações, como tamanho, grau, grau máximo, grau médio, diâmetro, distância média e variância de grau.

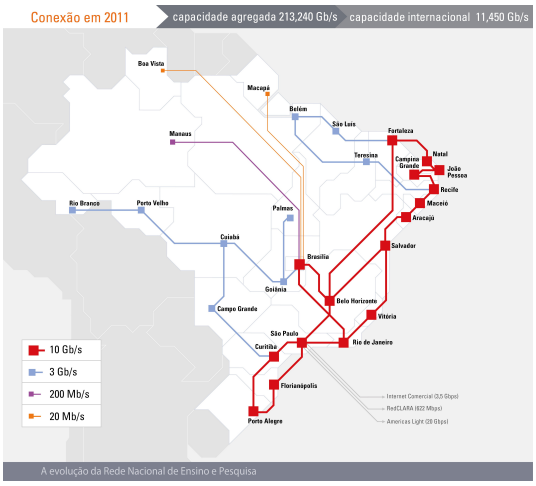
Sendo assim, este trabalho propõe implementar e testar os principais SGBD, como MariaDB (MARIADB, 2022), mySQL (MYSQL, 2022), MongoDB (MONGODB, 2022) e PostgreSQL (POSTGRESQL, 2022), para consultas costumeiras de bancos de grafos de redes de telecomunicações, analisar os resultados e definir o mais eficiente e adequado à situação.

2 JUSTIFICATIVAS

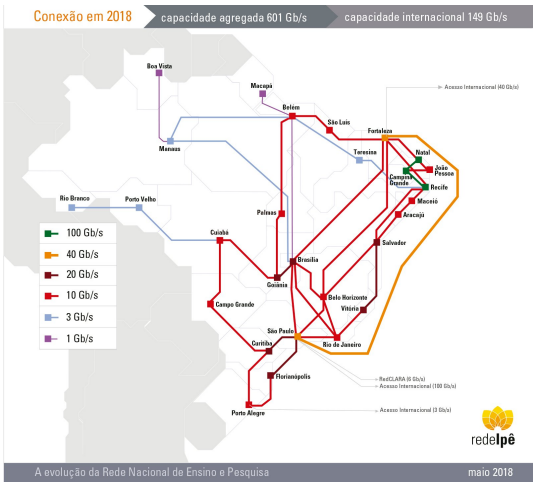
Com o crescimento e popularização da internet, que se tornou uma necessidade tanto industrial quanto doméstica, as malhas de redes de telecomunicações vêm se tornando cada vez mais extensas e complexas. Por exemplo, a RNP (Rede Nacional de Ensino e Pesquisa) aumentou a capacidade em 244% de 2010 para 2011, conforme ilustra a Figura 2(a), atingindo 213,2 Gb/s, e, em 2018, quase triplicou a capacidade, atingindo 601 Gb/s, conforme ilustra a Figura 2(b). A rede eduroam é parte da RNP e está disponível em universidades, centros de pesquisa, hospitais e centros públicos. Conta com mais de 3 mil pontos de acesso no Brasil e está presente em diversos países no mundo (RNP, 2022).

Figura 2 – Evolução da rede RNP

(a) Rede RNP em 2011



(b) Rede RNP em 2018



Fonte: RNP (2022)

O crescimento das redes de telecomunicações e seu consequente aumento de complexidade leva à necessidade de criação de sistemas de armazenamento de dados eficientes, capazes de oferecer rapidez e confiabilidade. O processamento destes dados só é possível se o resgate, registro e remoção das principais informações forem rápidos e eficientes. É no processo de extração de informações que os bancos de dados revelam-se fundamentais.

Ter conhecimento sobre o SGBD mais adequado a cada tipo de dado, portanto, é fundamental para garantir a rapidez e a efetividade dos sistemas. Neste trabalho, o objetivo é encontrar o SGBD com melhor desempenho para utilização nas diferentes consultas a bancos de dados de redes de telecomunicações com um grande volume de grafos relativamente pequenos (entre dez e vinte vértices).

3 OBJETIVOS

3.1 Objetivo Geral

O objetivo geral deste trabalho é encontrar o SGBD com maior desempenho para consultas de um banco de dados com uma quantidade volumosa de grafos.

3.2 Objetivos Específicos

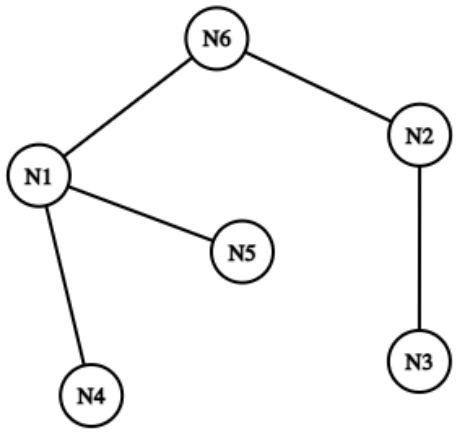
- Pesquisar e implementar *drivers* para diversos SGBD.
- Pesquisar e aplicar melhorias de desempenho para os *drivers* dos SGBD.
- Aferir e comparar o desempenho de diferentes SGBD.
- Estabelecer as vantagens e desvantagens de cada SGBD.
- Analisar os dados obtidos.

4 REFERENCIAL TEÓRICO

4.1 Teoria de Grafos

Um grafo G qualquer é composto de um conjunto de nós (também chamados de pontos ou vértices) ligados a outros nós por meio de arestas, que representam uma ligação atrelada a uma relação entre os nós (WEST, 2018). A Figura 3 ilustra um grafo de seis nós $N1, N2, N3, N4, N5, N6$ cujas ligações podem ser representadas pelos nós percententes àquela aresta, portanto, $N4N1, N1N5, N1N6, N6N2, N2N3$ são as arestas do grafo ilustrado.

Figura 3 – Exemplo de grafo com seis nós



Fonte: Produção do próprio autor.

Definições de acordo com Diestel (2017) de conceitos amplamentes utilizados em modelagens de redes de telecomunicações com grafos, ilustrados utilizando o grafo da Figura 3:

- **Ordem** : indica o número de vértices de um grafo. O grafo ilustrado possui ordem igual a 6.

- **Tamanho** : indica o número de arestas de um grafo. O grafo ilustrado possui tamanho igual a 5.
- **Grau** : o grau de um vértice é o número de arestas que o conectam. O grau do vértice $N1$ é 3, enquanto o grau do vértice $N4$ é 1.
- **Grau Máximo** : é o maior grau dos vértices do grafo. O grau máximo do grafo ilustrado é 3.
- **Grau Médio** : é a média aritmética dos graus de cada vértice. O grau médio do grafo ilustrado é 10/6.
- **Diâmetro** : é a maior distância entre quaisquer dois vértices. O diâmetro do grafo ilustrado é 4.
- **Distância Média** : é a média aritmética das distâncias de todos os pares de vértices de um grafo. A distância média do grafo ilustrado é 26/15.
- **Conectividade de Vértices** : é o menor número de vértices que podem ser removidos para desconectar o grafo. O grafo ilustrado possui conectividade igual a 1.
- **Variância de Grau**: é a variância de todos os graus dos vértices. O grafo ilustrado possui variância de grau igual a 2/3.

Na ciência da computação, grafos são utilizados em estruturas de dados para a representação de redes de telecomunicações, problemas de logística, *design* de circuitos elétricos, diagramas de árvore, árvore de decisões, algoritmo de Dijkstra, etc. Neste trabalho os dados de grafos representam várias pequenas redes de telecomunicações (entre dez a vinte vértices), com nós representando pontos fixos como roteadores, e suas ligações, representadas por arestas.

4.2 Bancos de Dados

De acordo com Silberschatz, Korth e Sudarshan (2011), bancos de dados são estruturas que contém dados inter-relacionados, tipicamente armazenados eletronicamente em sistemas de computadores, gerenciados por SGBD. Constituem parte essencial de qualquer comércio na atualidade, pois a todo momento usuários interagem com bancos de dados ao navegar na internet, mesmo que inconscientemente.

Os SGBD são úteis para proporcionar formas de armazenar e obter informações de bancos de forma simples e eficiente, ou seja, são responsáveis pelo gerenciamento dos bancos de

dados. SGBD permitem ao usuário criar e especificar esquemas (forma na qual os dados estão organizados) para os dados, fazer consultas ao banco, fazer consultas concorrentes aos mesmos dados, etc (GARCIA-MOLINA; ULLMAN; WIDOM, 2014). Em suma, provê ao usuário uma interface mais abstrata entre os dados e a aplicação.

Em 1970, Ted Codd, um matemático e pesquisador da IBM, propôs que os sistemas de bancos de dados apresentassem aos usuários dados organizados em forma de tabelas chamadas relações (CODD, 1970), que são conexões lógicas entre diferentes tabelas, baseadas em suas interações. Esses sistema é conhecido atualmente como bancos de dados relacionais. Este modelo contém uma ou mais tabelas com linhas e colunas; cada linha possui uma identificação (id) única e cada coluna representa um atributo. Também é possível associar linhas de diferentes tabelas gerando uma chave estrangeira.

A linguagem de programação dominante, utilizada por praticamente todos os bancos de dados relacionais, é a linguagem SQL (*Structured Query Language*), responsável pela administração de permissões, consulta e manipulação dos dados. Já bancos de dados não-relacionais (NoSQL) não necessitam de esquemas. Podem armazenar qualquer estrutura necessária, podendo alterá-la. Bancos de dados NoSQL possuem alta escalabilidade, isto é, aumentar seu volume de dados não impacta muito o desempenho. Apesar de serem menos escaláveis, os bancos de dados relacionais são melhores para tarefas que lidam com requerimentos complexos de relações entre os dados para modelação de sistemas igualmente complexos.

As principais diferenças entre os SGBD testados neste trabalho são quanto ao tipo, ou seja, relacional ou não-relacional, à capacidade e facilidade de escalabilidade, à licença de uso e aos tipos de dados suportados. Todos os SGBD testados possuem licença de uso compatível com o uso proposto.

5 METODOLOGIA E ETAPAS DE DESENVOLVIMENTO

5.1 Metodologia Adotada

De acordo com Gil (2002), este trabalho pode ser classificado como aplicado, por tentar solucionar um problema específico em uma circunstância particular. É classificado como descritivo, visto que seus objetivos são a coleta, análise e interpretação dos dados. Do ponto de vista dos procedimentos técnicos, é classificado como experimental, visto que, o trabalho busca identificar as variáveis do processo e suas dependências, e, mediante análise quantitativa, obter conclusões acerca dos dados coletados.

A princípio, serão feitos estudos de Teoria de Grafos, das diversas SGBD e suas peculiaridades e de métodos de aumento de desempenho em bancos de dados. Em seguida, serão escolhidos os SGBD mais interessantes e pertinentes ao uso de um grande volume de grafos relativamente pequenos. Após a escolha, serão realizados testes para avaliar o desempenho de cada SGBD a partir do tempo de consulta de operações comuns de grafos. Finalmente, será feita a análise dos dados coletados para levantamento de uma conclusão sobre o SGBD com melhor desempenho.

5.2 Cronograma de Trabalho

Apresenta-se nesta seção uma previsão do cronograma do plano de trabalho. O tempo total previso para a conclusão é de 5 meses.

Na Tabela 1, são detalhadas as atividades que se pretende realizar para o desenvolvimento do plano de trabalho. Assim, na primeira coluna da tabela são definidos os rótulos de cada atividade, e, na segunda coluna, é feita uma descrição da atividade que se pretende realizar. Finalmente, na Tabela 2, é apresentado o cronograma das atividades indicadas na Tabela 1.

Rótulo	Atividade
ATV 1	Estudo sobre grafos e bancos de dados e suas otimizações
ATV 2	Programação da comunicação com o SGBD
ATV 3	Testes de perfomance
ATV 4	Avaliação dos testes de perfomance
ATV 5	Escrita e revisão do projeto de graduação
ATV 6	Defesa do projeto de graduação

Tabela 1 – Lista de atividades.

Tabela 2 – Cronograma das atividades a serem efetuadas

Meses	Abril				Maio				Junho				Julho				Agosto			
Semanas	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Atividade 1																				
Atividade 2																				
Atividade 3																				
Atividade 4																				
Atividade 5																				
Atividade 6																				

Fonte: Produção do próprio autor.

6 ALOCAÇÃO DE RECURSOS

6.1 Recursos Materiais

O material bibliográfico utilizado é composto principalmente por periódicos científicos, artigos e livros. Este material estará disponível para o autor do trabalho das seguintes maneiras: fisicamente, via Biblioteca Central da UFES, e eletronicamente, via rede de internet da UFES, permitindo o acesso ao acervo eletrônico próprio da universidade e ao acervo cujo acesso tenha sido adquirido pela universidade.

6.2 Recursos Computacionais

As etapas de desenvolvimento do software serão realizadas utilizando a linguagem de programação *Python*, na sua versão mais atual, 3.10.2. Também será utilizada a tecnologia de conteinerização de código aberto *Docker*, para executar a aplicação num contêiner isolado. Serão testados os principais SGBD, como PostegreSQL, mySQL, mongoDB e MariaDB.

Os dados de grafos a serem utilizados nos testes foram gerados e adquiridos por Depizzol et al. (2018). São grafos que contém de dez a vinte nós, projetados para simular redes ópticas.

Além disso, o computador a ser utilizado nos experimentos é de propriedade do autor e possui a seguinte configuração: (*i*) sistema operacional Linux, distribuição Arch, Kernel versão 5.16; (*ii*) processador AMD 3600x, 3.80GHz com 6 núcleos físicos; (*iii*) memória RAM de 16 GB 3200MHz; (*iv*) armazenamento de 256GB (Unidade de estado sólido); (*v*) placa de vídeo Nvidia Geforce RTX 2060, com 6 GB de memória de vídeo dedicada.

REFERÊNCIAS

BIGGS E. KEITH LLOYD, R. J. W. N. Graph Teory 1736-1936. [S.l.]: Clarendon Press, 1986. Citado na página 8.

CARLSON, S. C. Königsberg bridge problem. Britannica, 2010. Disponível em: <https://www.britannica.com/science/Konigsberg-bridge-problem>. Acesso em: 24 mar. 2022. Citado na página 8.

CODD, E. F. A relational model of data for large shared data banks. Communications of the ACM, v. 13, n. 6, p. 377–387, 1970. Citado na página 14.

DEPIZZOL, D. B.; MONTALVãO, J.; LIMA, F. de O.; Moreira Paiva, M. H.; Vieira Segatto, M. E. Feature selection for optical network design via a new mutual information estimator. Expert Systems with Applications, v. 107, p. 72–88, 2018. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417418302483>. Citado na página 17.

DIESTEL, R. Graph theory. [S.l.]: Springer, 2017. Citado na página 12.

GARCIA-MOLINA, H.; ULLMAN, J. D.; WIDOM, J. Database systems: The complete book. 2. ed. [S.l.]: Pearson, 2014. Citado na página 14.

GIL, A. C. Como Elaborar Projetos de Pesquisa. 4. ed. [S.l.]: Atlas, 2002. Citado na página 15.

GUBICHEV, A. Query processing and optimization in graph databases. In: . [S.l.: s.n.], 2015. Citado na página 8.

MARIADB. MariaDB Foundation, 2022. Disponível em: <https://mariadb.org/>. Acesso em: 24 mar. 2022. Citado na página 9.

MONGODB. MongoDB, 2022. Disponível em: <https://mongodb.com>. Acesso em: 24 mar. 2022. Citado na página 9.

MYSQL. MySQL, 2022. Disponível em: <https://www.mysql.com>. Acesso em: 24 mar. 2022. Citado na página 9.

POSTGRESQL. PostreSQL, 2022. Disponível em: <https://www.postgresql.org>. Acesso em: 24 mar. 2022. Citado na página 9.

RNP. Rede nacional de ensino e pesquisa. In: _____. 2022. Disponível em: <https://www.rnp.br>. Acesso em: 20 mar. 2022. Citado na página 10.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. Database system concepts. [S.l.]: McGraw-Hill, 2011. Citado na página 13.

WEST, D. B. Introduction to graph theory. [S.l.]: Pearson, 2018. Citado na página 12.