UNIVERSIDADE FEEVALE

GUILHERME NUNES BEHS

ANÁLISE DE DESEMPENHO DE BANCOS NOSQL

Anteprojeto de Trabalho de Conclusão

Novo Hamburgo, março de 2020

GUILHERME NUNES BEHS

ANÁLISE DE DESEMPENHO DE BANCOS NOSQL

Anteprojeto de Trabalho de Conclusão de Curso, apresentado como requisito parcial à obtenção do grau de Bacharel em Ciência da Computação pela Universidade Feevale

Orientador: Juliano Varella de Carvalho

Novo Hamburgo, março de 2020

RESUMO

O conceito de *Big Data* tem se mostrado muito presente no cotidiano atual, desde o momento em que empresas e órgãos públicos buscaram as melhores técnicas e ferramentas para ter *highlights* do passado e antecipar ações futuras com grandes volumes de dados. Estes dados, muitas vezes desestruturados e de várias fontes, vêm sendo melhor aproveitados a fim de localizar padrões e tendências que podem ser trabalhadas desde o momento em que são descobertas. Entre técnicas de visualização e algoritmos de previsão, ainda existe o processo de armazenamento e recuperação destes dados, que varia dependendo da estrutura dos mesmos. Bancos relacionais e não-relacionais têm estado no meio desta discussão, principalmente os não-relacionais, também conhecidos como NoSQL (*Not Only SQL*). Estes se destacam pela agilidade em consultas oriundas da abdicação às regras de integridade e relacionamento. Dados públicos são de extrema importância para a sociedade, a fim de que se possa entender e estudar o que está acontecendo, e aplicar as devidas medidas. Os dados do ENEM são públicos e podem ser estudados com o propósito de analisar a educação brasileira. Este trabalho investigará o banco de dados não-relacional mais apropriado para armazenar e consultar os *datasets* do ENEM, dentre um conjunto de SGBDs (Sistema Gerenciador de Bancos de Dados) previamente selecionados, estudando as melhores práticas de consulta em cada um deles e usando ferramentas para monitorar seus desempenhos, a fim de fazer comparações e obter uma conclusão mais precisa.

Palavras-chave: Big Data .NoSQL. banco de dados. dados públicos. enem.

SUMÁRIO

MOTIVAÇÃO

Devido à geração massiva de dados, cada vez mais frequente, surgiu a necessidade de tecnologias diversas trabalharem com *Big Data*. Cave et al. (2020) definiram Big Data como sendo uma extrema quantidade de dados que pode ser complexa, multidimensional, desestruturada e heterogênea. Estes dados podem ser estudados a fim de buscar padrões e estatísticas para o campo em que estão inseridos.

Segundo o Canaltech (2020), *Big Data* engloba o conceito dos 5Vs: Volume, Variedade, Velocidade, Veracidade e Valor. Os três primeiros Vs se referem a grandes quantidades de dados não-estruturados e que precisam ser analisados rapidamente. **Veracidade** se refere à confiança da fonte e qualidade dos dados. Já **Valor** se refere ao custo-benefício da análise destes dados e de que forma eles podem beneficiar os negócios.

O Domo(2020) mostrou algumas estatísticas em relação a geração de dados em 2019. Segundo ele, por minuto, usuários do Twitter enviaram 511.200 tweets, usuários do Instagram postaram 55.140 fotos, passageiros do Uber realizaram 9.772 corridas, e o Google realizou 4.497.420 pesquisas. Estes dados enfatizam o crescimento de dados em um curto prazo de tempo, e alertam para a urgência no desenvolvimento de ferramentas cada vez mais eficientes para a manipulação e consulta destes dados.

Como exemplos de *Big Data* e aplicações, é possível citar dados eletrônicos bancários, onde sua análise pode determinar futuros empréstimos e concessão de benefícios por parte do banco. Serviços de *E-Commerce*, como a Amazon, e de *Streaming*, como Spotify e Netflix, também se beneficiam de *Big Data* utilizando o histórico de pesquisa e consumo de seus usuários para fazer sugestões de consumo.

Xiao, Silva e Zhang (2020) abordaram um estudo em Shangai, na China, para avaliar o equilíbrio entre casa e trabalho de pessoas que vivem sob o sistema "996", onde as pessoas trabalham das 9h da manhã até ás 9h da noite por 6 dias na semana, e o excesso de horas extras. A avaliação foi efetuada utilizando grandes volumes de dados provenientes do GPS do celular, para identificar pontos de circulação destas pessoas, e quanto tempo elas ficam nestes pontos.

Finkenstein et al. (2020) utilizaram técnicas de *Big Data* para estudar e identificar as melhores práticas para aprimorar o fornecimento de assistência médica dentária com estudos aplicados e comprovados, baseando-se em conjuntos de dados fornecidos por entidades de saúde diversas. Este trabalho e o de Xiao, Silva e Zhang (2020) mostram que *Big Data* não se restringe ao campo financeiro e comercial, mas também na análise de comportamento das pessoas e na melhoria de serviços.

Big Data uniu profissionais de banco de dados e cientistas de dados. Ambos profissionais têm buscado formas mais eficazes de manipular tais dados, para que sua recuperação não se torne um obstáculo na hora de analisá-los.

O trabalho de Freire et al. (2016) fez um estudo guiado em relação ao desempenho entre SGBDs (Sistema Gerenciador de Bancos de Dados) de diferentes estruturas que continham, cada um, grandes volumes de registros médicos eletrônicos provenientes de diversas fontes. Este estudo destacou a importância da estrutura dos bancos de dados, entre modelos relacionais e não-relacionais, para a performance das consultas.

Os bancos de dados relacionais atuais, como MySQL e Oracle, se destacam pelas propriedades ACID (Atomicidade, Consistência, Integridade, Durabilidade). Essas propriedades ligadas aos modelos relacionais foram, por muito tempo, seus principais motivos de uso. No entanto, como dito por Chen e Lee (2019), as inúmeras relações entre suas tabelas podem atrasar a recuperação de seus dados.

Em meio a tal demanda, bancos de dados NoSQL (*Not Only SQL*) começaram a se destacar devido a sua arquitetura simples e que não enfatiza a importância das propriedades ACID. Sua performance no retorno dos dados em contrapartida a bancos relacionais se mostrou mais eficaz em alguns casos de Big Data, como os testes realizados por Freire et al. (2016). Chen e Lee (2019) acrescentam que o modelo NoSQL também visa o escalonamento horizontal de estruturas de dados heterogêneas e um suporte simples para replicação *master-slave* e *peer-to-peer*.

O trabalho de McDonald et al. (2019) envolveram o desenvolvimento do redbiom, um sistema que identifica e caracteriza comunidades microbianas. A necessidade deste sistema se deu pela alta quantidade de amostras disponíveis e metadados para fazer a análise, que tomavam muito tempo dos pesquisadores. Armazenando estes dados no Redis, um banco

de dado NoSQL *in-memory*, foi possível agilizar o processo de busca das amostras necessárias.

Os atuais bancos de dados NoSQL podem ter sua estrutura interna em colunas (Cassandra, HBase), documentos (MongoDB, CouchDB), chaves-valor (Redis, Dynamo), grafos (OrientDB, Neo4J), entre outros. Todos com sua devida aplicação, variando na estrutura e complexidade de seus dados

A manipulação de um grande volume de dados públicos pode ser uma das principais ferramentas para gestores públicos aplicarem as devidas medidas nos setores mais necessitados. Porém, a má compreensão de como aplicar técnicas de *Big Data* pode impedir este ato. O estudo de Guenduez, Mettler e Schedler (2020) ressaltou o fato de que gestores públicos não tem total compreensão, nem confiança nas atuais técnicas de Big Data aplicadas aos dados públicos. Isso reforça ainda mais a urgência por trabalhos na área que envolvam dados públicos.

O volume de dados armazenados em função do ENEM (Exame Nacional do Ensino Médio) é uma boa fonte de análise e estudo. Seus dados são públicos e armazenados em documentos CSV e, atualmente, ultrapassam 20 GB. Há dados desde a primeira aplicação do ENEM (1998) até o ano de 2018, no momento da escrita deste trabalho. Estudá-los e manipulá-los é de suma importância para a compreensão de diversos aspectos educacionais brasileiros

Vale salientar que esse trabalho se integra com o trabalho de conclusão do aluno Gustavo Willrich, também da Universidade Feevale, que desenvolverá uma plataforma de visualização de dados com a base do ENEM.

Este trabalho, portanto, propõe a análise de desempenho de um conjunto de SGBDs NoSQL. Utilizando o *dataset* público do ENEM, e comunicação constante com o aluno Gustavo Willrich, será avaliado o SGBD mais adequado para a manipulação dos dados do ENEM.

OBJETIVOS

Objetivo Geral:

Avaliar o desempenho de bancos de dados NoSQL, utilizando dados públicos abertos.

Objetivos Específicos:

- Investigar características de bancos de dados NoSQL.
- Selecionar três bancos de dados NoSQL.
- Caracterizar os três bancos de dados NoSQL escolhidos.
- Criar a infraestrutura necessária para utilização dos bancos de dados.
- Conhecer detalhadamente o *dataset* a ser utilizado.
- Definir consultas simples e complexas em cada um dos SGBDs escolhidos.
- Identificar o melhor SGBD para cada tipo de consulta efetuada.

METODOLOGIA

Neste trabalho será realizada uma pesquisa de natureza aplicada. Para isto, será feita uma avaliação de desempenho de consultas sob uma amostragem inserida em um conjunto de bancos de dados NoSQL, cada um com a mesma quantidade de registros provenientes dos documentos do ENEM, cujo download será efetuado no site http://inep.gov.br/microdados .

Os dados fornecidos pelo ENEM serão estudados para entender melhor a relação deles entre si, suas particularidades e as informações contidas nos metadados. Após o estudo, haverá uma integração desta pesquisa com o trabalho do aluno Gustavo Willrich, a fim de entender as necessidades de buscas de dados sobre seu trabalho.

Os bancos de dados NoSQL selecionados serão estudados detalhadamente, suas estruturas, formas de armazenamento, técnicas de consulta e índices. Baseado nas discussões com o aluno Gustavo Willrich, serão elaboradas consultas que pontuam as necessidades do trabalho, estudando índices e organizações nas coleções dos SGBDs.

A avaliação de desempenho irá monitorar recursos do terminal que executará as consultas com ferramentas destinadas a esse propósito. Para evitar a influência de *frameworks* e *middlewares*, as consultas serão executadas diretamente em *terminal clients* de cada SGBD.

Para a avaliação de desempenho, serão analisados o espaço em disco ocupado pelos arquivos e os SGBDs, e o tempo de resposta em relação às consultas. As consultas serão medidas em uma tabela de comparações que exibirá a média, o mínimo, o máximo, o desvio padrão, o menor quartil, o maior quartil, e a mediana do tempo.

Seguindo esta metodologia e baseando-se nos resultados obtidos, será possível responder a seguinte questão: Qual é o banco de dados não-relacional mais adequado para o armazenamento dos *datasets* do ENEM?

CRONOGRAMA

Trabalho de Conclusão I

Etapa	Meses			
	Mar	Abr	Mai	Jun
Anteprojeto				
Pesquisar SGBDs NoSQL apropriados para o trabalho				
Estudar os SGBDs selecionados				
Pesquisar ferramentas para medir desempenho das consultas				
Estudar os Dados do ENEM				
Elaborar TC I				

Trabalho de Conclusão II

Etapa	Meses			
	Ago	Set	Out	Nov
Inserir amostragem de dados nos SGBDs				
Conversar com o aluno Gustavo Willrich				
Elaborar as consultas				

Executar e medir o desempenho das consultas				
Elaborar TC II				

BIBLIOGRAFIA

CAVE, Alison et al.. **Big Data – How to Realize the Promise.** Clinical Pharmacology & Therapeutics, *107*(4), 753–76, abril de 2020. Disponível em <https://doi.org/10.1002/cpt.1736>. Acesso em 22 de Março de 2020.

CHEN, jean-kuo; LEE, wei-zhe. **An Introduction of NoSQL Databases Based on Their Categories and Application Industries**. Algorithms, 12, 106, maio de 2019. Disponível em <https://doi.org/10.3390/a12050106>. Acesso em 22 de Março de 2020.

DOMO. **Data Never Sleeps 7.0** . Domo. Disponível em <https://www.domo.com/learn/data-never-sleeps-7>. Acesso em 28/03/2020.

FREIRE, Sergio et al.. **Comparing the Performance of NoSQL Approaches for Managing Archetype-Based Electronic Health Record Data**. PLOS ONE. 11. e0150069, março de 2016. Disponível em <https://doi.org/10.1371/journal.pone.0150069>. Acesso em 22 de Março de 2020.

FINKELSTEIN, Joseph et al.. **Using big data to promote precision oral health in the context of a learning healthcare system**. Journal of Public Health Dentistry, 80(S1), S43–S58, março de 2019. Disponível em <https://doi.org/10.1111/jphd.12354>. Acesso em 8 de Março de 2020.

GUENDUEZ, Ali A.; METTLER, Tobias; SCHEDLER, Kuno. **Technological frames in public administration: What do public managers think of big data?** Government Information Quarterly, 37(1), janeiro de 2020. Disponível em <https://doi.org/10.1016/j.giq.2019.101406>. Acesso em 22 de Março de 2020.

MCDONALD, Daniel et al.. (2019). **redbiom: a Rapid Sample Discovery and Feature Characterization System** OBSERVATION Novel Systems Biology Techniques, junho de 2019. Disponível em <https://msystems.asm.org/content/4/4/e00215-19>. Acesso em 8 de Março de 2020.

CANALTECH. O que é Big Data? . **Canaltech**. Disponível em <https://canaltech.com.br/big-data/o-que-e-big-data/>. Acesso em 28 de Março de 2020.

XIAO, Chaowei; SILVA, Elisabete A.; ZHANG, Chuchu. **Nine-nine-six work system and people's movement patterns: Using big data sets to analyse overtime working in Shanghai**. Land Use Policy, 90, janeiro de 2020. Disponível em <https://doi.org/10.1016/j.landusepol.2019.104340>. Acesso em 22 de Março de 2020.