UNIVERSIDADE FEDERAL DO AMAPÁ
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO
DEPARTAMENTO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

AVALIAÇÃO DO BANCO DE DADOS NOSQL HBASE NO CONTEXTO DE DADOS GEORREFERENCIADOS

RODRIGO SANTOS BALIEIRO

Orientador: Thiago Pinheiro do Nascimento

MACAPÁ
FEVEREIRO DE 2019

RODRIGO SANTOS BALIEIRO

AVALIAÇÃO DO BANCO DE DADOS NOSQL HBASE NO CONTEXTO DE DADOS GEORREFERENCIADOS

Trabalho de Conclusão de Curso apresentado à Universidade Federal do Amapá como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Thiago Pinheiro do Nascimento

MACAPÁ
FEVEREIRO DE 2019

RODRIGO SANTOS BALIEIRO

AVALIAÇÃO DO BANCO DE DADOS NOSQL HBASE NO CONTEXTO DE DADOS GEORREFERENCIADOS

Trabalho de Conclusão de Curso apresentado à Universidade Federal do Amapá como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Trabalho aprovado. Macapá, de de 2018.

Thiago Pinheiro do Nascimento
Orientador

Dr. José Walter Cárdenas Sotil
Universidade Federal do Amapá

Me. Marco Antônio Leal da Silva
Universidade Federal do Amapá

MACAPÁ
FEVEREIRO DE 2019

Dedico este trabalho a comunidade acadêmica de Ciência da Computação, a qual deixo esta singela contribuição bibliográfica.

Em especial a minha esposa Aline e meu filho Paulo Henrique.

Agradecimentos

Inicialmente, agradeço à Deus, pela vida, saúde e livramentos que até aqui me mantiveram para o desenvolvimento deste trabalho...

A Família pelo apoio, carinho e sacrifício nas diversas noites em que me ausentei para as pesquisas...

Ao Orientador por abraçar a causa e dedicar-se junto a mim nos mais diversos desafios que foram proporcionados pelo tema...

Aos Amigos que sempre estiveram presentes nas adversidades e alegrias...

À Universidade Federal do Amapá – UNIFAP, ...

Finalmente, agradeço à todos que contribuíram de forma direta ou indireta para a construção desse trabalho.

"O perigo de verdade não é que computadores passem a pensar como humanos, mas sim que humanos passem a pensar como computadores".
– Sydney Harris

Resumo

Aplicações de georreferenciamento em tempo real usualmente armazenam e gerenciam grandes quantidades de dados produzidos em baixa latência de tempo. Essa situação é desafiadora quando essas aplicações são orientadas a banco de dados relacionais, que possuem baixa escalabilidade para armazenarem e gerenciarem excessivos volumes de dados voláteis e não-voláteis. Este trabalho propõe uma alternativa para esse contexto.

Palavras-chave: Georreferenciamento. Escalabilidade. Banco de Dados.

Abstract

Application of real time georeferencing are traditionally stored and managed in large amountsofdatainlowlatencytime.Thissessioniseffectivewhentheseapplicationsare oriented to relational databases, which have low scalability to store and manage large volumes of volatile and non volatile data. This work is an alternative to this context.

Keywords: Georeferencing. Scalability. Database.

Lista de Figuras

Lista de Abreviaturas e Siglas

IBM	International Business Machines - Empresa Multinacional de Informática.
IOT	Internet Of Things - Internet das Coisas.
MQTT	Message Queuing Telemetry Transport - Protocolo de Telemetria e transporte de mensagens.
API	Application Programming Interface - Interface de Programação de Aplicativos.
Javascript	Linguagem de Programação fracamente tipada voltada para Sistemas Web.
WEB	Sigla em inglês para referir-se a rede mundial de computadores.
SQL	Service Query Langage - Linguagem Específica de Programação para Banco de Dados

Sumário

1 Introdução

O presente trabalho aborda o contexto dos sistemas de georreferenciamento de tempo real, que, usualmente necessitam de abordagens escaláveis para o gerenciamento e o armazenamento de grandes quantidades de dados, mas que tradicionalmente são implementados para funcionarem sobre sistemas de bancos de dados relacionais, que dispõem de pouca escalabilidade para armazenar e gerenciar excessivos volumes de dados produzidos em baixa latência de tempo. Para contornar essa situação, esta pesquisa visa apresentar uma solução viável para essa problemática, onde ela será utilizada para a geolocalização de viatura da Guarda Civil Municipal de Macapá. Nesse trabalho, os aspectos essenciais da pesquisa proposta são apresentados e discutidos, tais como: a problemática a ser solucionada, os objetivos e justificativas, a metodologia e o cronograma de atividades a serem realizadas e desenvolvidas. Ao final desse capítulo introdutório, também é apresentada a organização do trabalho proposto.

1.1 Contexto

De acordo com uma pesquisa realizada pelo IBGE, cada vez mais residências no Brasil têm computadores ou outros dispositivos conectados à internet[1]. Isto mostra que as informações digitais estão cada vez mais presentes no cotidiano da sociedade e não apenas nos negócios.

Para Kevin Ashton, as pessoas costumavam guardar seus objetos em gavetas de armários, dentro de cômodas, cofres e até debaixo da cama, depois da digitalização deles, passaram a guardar em banco de dados [2].

Os Banco de Dados foram criados para guardar informações. Elas podem ser de diferentes tamanhos e formatos que ficam armazenados em espaços organizados e categorizados[3].

A organização dos Dados dentro de um Banco de Dados pode variar, mas geralmente, possui quatro operações básicas, dentre elas: adicionar, remover, copiar e atualizar os dados. Essas operações geralmente são realizadas por um software que gerencia e controla todo o fluxo que se passa em um Banco de Dados. Esses softwares são chamados de Sistemas de Gerenciamento de Banco de Dados – SGBD[4].

Inicialmente os dados computacionais eram armazenados em arquivos[3], o que dificultava seu gerenciamento porque todas as informações eram postas de forma desorganizada e não categorizada, fazendo com que uma busca, por exemplo, fosse uma tarefa complicada. Além das informações estarem desorganizadas, ainda havia o

problema da confiabilidade e a durabilidade dos dados armazenados, comprometendo, de forma geral, a segurança das informações, pois quem tivesse acesso ao computador onde os dados fossem armazenados, poderia manipular o arquivo, comprometendo tudo e, ainda, qualquer pane que o computador apresentasse iria influenciar no registro de informações[3].

Posteriormente, as informações foram melhor organizadas em colunas e linhas, formando tabelas, conhecidas como relações, o que permitiu muitas soluções a problemas de armazenamento para as corporações e para os sistemas gerenciadores de banco de dados[4].

Durante a década de 1960 vários estudos foram realizados para formação de diferentes modelos de Banco de Dados, como o Hierárquico e o de Redes. Entretanto, foi a partir da década seguinte, em 1970, com a publicação do artigo *A Relational Model of Data for Large Shared Data Banks*, de Edgar Frank Codd, que foi apresentado o modelo Relacional[4].

O modelo relacional surgiu para aumentar a independência dos Dados nos SGBDs e para compor uma estrutura mais sólida e durável para o armazenamento e recuperação dos dados [3]. Com esse modelo, algumas características que outros modelos não possuíam surgiram, como não ter caminhos pré-definidos para se fazer acesso aos dados; implementar estruturas de dados organizadas em relações e; restrição de repetição de informações [3].

Com a popularização e aperfeiçoamento do modelo relacional e utilização da linguagem SQL pelos Sistemas de Gerenciamento de Banco de Dados, por muitos anos as corporações e até mesmo donos de sites utilizaram os Banco de Dados Relacionais como armazenamento de suas informações[5]. Como os dispositivos conectados a internet passaram a ser domésticos, o fluxo de dados aumentou exponencialmente, exigindo tanto uma maior capacidade e velocidade da Internet quanto do armazenamento digital[5].

Tendo em vista número crescente de dispositivos e pessoas conectados à internet [1], muitos órgãos governamentais e não-governamentais têm utilizado sistemas informatizados para obter informações relevantes sobre dados dos mais variados tipos para que se possam obter, a partir desses dados, novas informações específicas, como estatísticas, variações de velocidade, taxas de consumo, mapeamento de regiões, dentre outros que podem estar em um universo comumente conhecido como Internet das Coisas, do termo original, em inglês, *Internet Of Things*[6].

Muitos sistemas de georreferenciamento de tempo real utilizam os conceitos de *"Internet of Things"*, de forma a possibilitar que aparelhos eletrônicos compartilhem dados de localização [7]. Esse compartilhamento de dados georreferenciados são utilizados em diversos domínios de aplicações, que vão desde as situações de desastes

naturais, onde usuários envolvidos em operações de proteção civil necessitam de graus eficientes e eficazes de coordenação [8]; até para medição remota, armazenamento, processamento e visualização de informações químicas [7].

O pesquisador Cézar Taurion da Universidade de São Paulo explica que todos os dados gerados por grandes empresas, redes sociais, sensores e dispositivos de diferentes formas formam volumes que ultrapassam a unidade dos zetabytes ou 10^{21} bytes[5], compondo o que é conhecido como *Big Data*[5].

Taurion ainda apresenta dois conceitos de *Big Data*. O primeiro se refere a um conjunto de dados cujo crescimento é exponencial e cuja dimensão está além da habilidade das ferramentas típicas de capturar, gerenciar e analisar dados[5]. O segundo diz que *Big Data* é o termo usado pelo mercado para descrever problemas no gerenciamento de processamento de informações em grandes volumes as quais ultrapassam a capacidade das tecnologias de informações tradicionais ao longo de uma ou mais dimensões[5].

Como os dados têm aumentado de forma explosiva devido ao desenvolvimento de redes sociais e computação em nuvem, tem havido um novo desafio para armazenar, processar e analisar um grande volume de dados. As tecnologias tradicionais não se tornam uma solução adequada para processar big data, de modo que uma plataforma de big data começou a surgir[9].

Para esses problemas citados, tecnologias foram desenvolvidas especificamente para tratar grandes volumes de informações e diversificadas fontes de geração de tais dados. Uma solução seria a adoção da linguagem NoSQL (Not Only SQL), onde a construção das tabelas é diferente da tradicional e formada para os mais variados tipos; Outro método adotado seria o uso dos Banco de Dados Não Relacionais; Uma outra abordagem seria os Sistemas em Nuvem[5].

1.2 Problemática

(Sistemas de georeferenciamento armazenam grandes volumes de dados, os quais os bancos de dados não relacionais não conseguem suportar).

Como este trabalho conterá experimentação de aplicações que envolvem Banco de Dados, *Big data* e Georreferenciamento é importante destacar que:

Na sociedade atual são crescentes os dispositivos que se conectam à internet e dispõem informações que podem ser usadas e podem ser aproveitadas para diversas aplicações computacionais[7]. Dentre elas existem as que apresentam informações e cálculos de lugares, distâncias, clima, altura, velocidade, em que é muito comum o desenvolvimento de Sistemas de Informações Geográficos (SIG's) [7]. Esses siste-

mas são capazes de armazenar dados cartográficos, censitários, cadastros urbano e rural e imagens de satélite, dispondo de mecanismos para tratar as informações, assim como para consultar, recuperar, visualizar e plotar o material na base de dados georreferenciados[7].

Geralmente os dados gerados por Sistemas de Informações Geográficas ou Geoespaciais são dinâmicos, pois para se obter precisão é necessário que esses dados sejam atualizados em tempo real, acumulando um grande volume de dados para serem armazenados em um banco de dados do tipo relacional[10].

Por esse motivo, faz-se necessário experimentar outro tipo de banco de dados, em que seja suportado um grande fluxo, com maior escalabilidade e tempo de resposta na transição de informações georreferenciais.

1.3 Objetivos

Os objetivos do presente trabalho de pesquisa são apresentados em torno de um único objetivo geral e três objetivos específicos.

1.3.1 Objetivo Geral

Avaliar o sistema de bancos de dados não relacional orientado a coluna HBase no armazenamento e no gerenciamento de volumes de dados georreferenciados de tempo real, no contexto da geolocalização de viaturas da Guarda Municipal de Macapá.

1.3.2 Objetivos Específicos

O presente trabalho de pesquisa tem os seguintes objetivos específicos:

- Estudar o contexto dos sistemas de geolocalização em tempo real.
- Avaliar sistemas de banco de dados escaláveis para grandes volumes de dados.
- Propor um estudo de caso para avaliar e comparar banco de dados relacionais e banco de dados não-relacional HBase para armazenarem e gerenciarem excessivos volumes de dados georreferenciados de tempo real, produzidos em baixa latência.

1.4 Hipótese

Ao se utilizar dos recursos de software para *Big Data*, como supracitado na contextualização por Oussous [11], existe a possibilidade de analisar o desempenho de uma aplicação baseada em georreferenciamento de tempo real quando aplicada ao cenário de Segurança Pública.

1.5 Justificativa

Com a popularização e aumento do uso de aparelhos que são detectados por localização como é o caso de telefones celulares, GPS e centrais automotivas, por exemplo, os dados de localização aumentaram consideravelmente nas últimas décadas do século XXI, o que é um desafio para o armazenamento tradicional de dados digitais[12].

Para estudiosos como Shouwu He, Longxian Chu e Xiauying Li, existem formas de lidar com os desafios de se armazenar, gerenciar e acessar grandes volumes de dados originados por diversas fontes diferentes. Uma das soluções seria com a utilização do banco de dados HBase[12].

Já para Jian Huang *Et al*[13], O banco de dados HBase é um sistema de armazenamento chave/valor de código livre, então, de fácil acesso, que está sendo usado em *datacenters* de grandes empresas como Facebook e Twitter devido sua portabilidade e escalabilidade massiva.

Se o Habse for usado na saúde, de acordo com Yang jun *Et al*[14], como o foco para pesquisas e cuidados da saúde se baseia na análise de estatísticas e uso de dados em situações posteriores, o armazenamento massivo bem como a eficácia do sistema *Hadoop Distributed File System* – HDFS é a chave para se obter resultados esperados e, até mesmo, diagnósticos médicos.

1.6 Metodologia

Para o desenvolvimento e experimentação do trabalho proposto serão utilizados procedimentos metodológicos divididos em duas etapas. O referencial bibliográfico e a experimentação científica.

O referencial bibliográfico destaca não somente a pesquisa textual, mas a base para argumentação, além do desenvolvimento dos códigos a serem usados nos aplicativos dos testes computacionais.

Para isto será utilizada a revisão de literatura *Ad Hoc* que é um método em que a revisão é feita de forma livre, sem o uso de guias ou fases a serem seguidas[15].

Já a experimentação científica tem como proposta os testes e a coleta de dados para que depois sejam levantadas conclusões com base em análises objetivas ou subjetivas feitas por meio de números ou percentuais.

Desta forma o método a ser usado será o Estudo de Caso, em que um determinado fenômeno será analisado, em seu ambiente e curso naturais depois da utilização de um software[15].

E ainda, vale destacar a utilização do método de "Metanálise", pois a partir

dele é possível fazer estudos que integram resultados sobre determinada questão pesquisada, combinando e fazendo comparativos entre eles[15]. Esse método será usado, principalmente no comparativo dos Sistemas de Gerenciamento de Banco de Dados (SGBDs), em que diferentes softwares serão testados, como *MySQL*, *PostgreeSQL* e a plataforma *Apache Hadoop*.

«««««««««««»»»»»»»»»»»» Esta parte aqui debaixo será analisada pelo professor Thiago, pois pode não ser necessária a inclusão dela neste momento.

Como base para o trabalho foi utilizado o artigo, da empresa IBM[16], que é um tutorial para o desenvolvimento de um aplicativo para celular. Tal aplicação usa o GPS do aparelho para enviar informações de geolocalização (latitudes e longitudes) para um ambiente de nuvem da própria IBM.

Esses dados podem ser consumidos por outra ferramenta digital, neste caso, seguindo o artigo de tutorial, os dados foram aproveitados por um sistema *web* escrito em linguagem PHP, usando o protocolo MQTT, para intercomunicação entre as aplicações do hardware do celular, da nuvem e do computador de visualização. Esse esquema pode ser melhor visualizado na Lista de Figuras.

Maiores detalhes desta abordagem serão apresentados nos próximos capítulos.

Vale salientar que, embora haja neste trabalho uma contribuição de escrita no código e modificações nos aplicativos seguidos no tutorial da IBM, o objetivo central aqui proposto não está focado no desenvolvimento da ferramenta, mas na utilidade de sua aplicabilidade. Sendo assim, os resultados a serem alcançados tentarão satisfazer as hipóteses levantadas.

Referências

[1] E. Sociais, "Pnad contínua tic 2017: Internet chega a três em cada quatro domicílios do país," *Agencia de Notícias - IBGE*.

[2] K. Ashton, "Beginning the internet of things," *Medium*, p. 1, 2016.

[3] O. K. Takai, I. cristina Italiano, and J. E. Ferreira, "Introdução a banco de dados," *IME-USP*, vol. único, p. 124, 2005.

[4] C. Date, *Introdução a Sistemas de Bancos de Dados*, vol. 1. 7 ed.

[5] C. Taurion, *Big Data*, vol. 1. 1 ed.

[6] E. Magrani, *A Internet Das Coisas*, vol. único. 1ª ed.

[7] F. Vega, J. Pantoja, S. Morales, O. Urbano, A. Arevalo, E. Muskus, C. Pedraza, M. Patino, M. Suarez, and N. Hernandez, "An iot-based open platform for monitoring non-ionizing radiation levels in colombia," in *2016 IEEE Colombian Conference on Communications and Computing (COLCOM)*, pp. 1–4, April 2016.

[8] T. Barroso, J. Sanguino, and A. Rodrigues, "Georeferencing for coordinated positioning applications," in *2011 The 14th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, pp. 1–5, Oct 2011.

[9] K. Park, M. C. Nguyen, and H. Won, "Web-based collaborative big data analytics on big data as a service platform," in *2015 17th International Conference on Advanced Communication Technology (ICACT)*, pp. 564–567, July 2015.

[10] P. Parikh and T. D. Nielsen, "Iee colloquium on 'experience in the use of geographic information systems in the electricity supply industry (digest no.129)," in *IEE Colloquium on Experience in the Use of Geographic Information Systems in the Electricity Supply Industry*, pp. 1–5, Oct 2011.

[11] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "Big data technologies: A survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 4, pp. 431 – 448, 2018.

[12] S. He, L. Chu, and X. Li, "Spatial query processing for location based application on hbase," in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(*, pp. 110–114, March 2017.

[13] J. Huang, X. Ouyang, J. Jose, M. Wasi-ur-Rahman, H. Wang, M. Luo, H. Subramoni, C. Murthy, and D. K. Panda, "High-performance design of hbase with rdma over infiniband," in *2012 IEEE 26th International Parallel and Distributed Processing Symposium*, pp. 774–785, May 2012.

[14] Y. Jin, T. Deyu, and Z. Yi, "A distributed storage model for ehr based on hbase," in *2011 International Conference on Information Management, Innovation Management and Industrial Engineering*, vol. 2, pp. 369–372, Nov 2011.

[15] P. Jacobsen, "Como escolher o método de pesquisa."

[16] V. Vaswani, "Desenvolva um aplicativo em php que use dados de gps de um dispositivo iot."