

Análise Comparativa de Desempenho em Drivers para MongoDB em Aplicações Node.js

Leandro Ungari Cayres, Ronaldo Celso Messias Correia

¹Faculdade de Ciências e Tecnologia – Universidade Estadual Paulista (UNESP)
Presidente Prudente – SP – Brazil

{leandro.ungari, ronaldo.correia}@unesp.br

Abstract. *This meta-paper describes the style to be used in articles and short papers for SBC conferences. For papers in English, you should add just an abstract while for the papers in Portuguese, we also ask for an abstract in Portuguese (“resumo”). In both cases, abstracts should not have more than 10 lines and must be in the first page of the paper.*

Resumo. *Este meta-artigo descreve o estilo a ser usado na confecção de artigos e resumos de artigos para publicação nos anais das conferências organizadas pela SBC. É solicitada a escrita de resumo e abstract apenas para os artigos escritos em português. Artigos em inglês deverão apresentar apenas abstract. Nos dois casos, o autor deve tomar cuidado para que o resumo (e o abstract) não ultrapassem 10 linhas cada, sendo que ambos devem estar na primeira página do artigo.*

1. Introdução

Nos últimos anos, o crescimento no volume de dados mudou a utilização desses por empresas e organizações; tais dados, inicialmente considerados agentes passivos, relacionados às regras de negócio empresarial; tornaram-se potenciais oportunidades de lucro através da análise de informações, e consequentemente, do conhecimento presente no conjunto de dados.

Essa crescente quantidade de dados, chamado de *Big Data*, não somente requer maior espaço de armazenamento, mas uma mudança em sua organização com base em cada contexto, considerando características como volume, variedade, velocidade e valores [Ward and Barker 2013]. A arquitetura dos tradicionais bancos de dados relacionais, baseada no modelo ACID (*atomicity, consistency, isolation e durability*), contudo, em ambientes de *Big Data*, a alta consistência afeta diretamente os aspectos de disponibilidade e eficiência, que são importantes, devido ao alto volume, variedade e velocidade presente em *Big Data* [González-Aparicio et al. 2016].

Nesse cenário surgem os banco de dados NoSQL (“Not only SQL”) provendo maior flexibilidade estrutural, suporte a replicação e consistência eventual seguindo o critério BASE (*basic, availability, soft-state e eventual consistency*) [Han et al. 2011]. Nos últimos anos, a popularidade dos banco de dados não-relacionais tem crescido [Cooper et al. 2010, Edlich 2015] proporcionando diversas soluções conforme características dos dados cada aplicação.

Diferentemente do modelo relacional, os bancos NoSQL não possuem uma linguagem comum entre eles que permita a realização de operações, como o SQL (*Structured*

Query Language), desse modo, cada banco provê uma interface nativa para manipulação dos dados. Contudo, uso de chamadas de sistema para execução de comandos dessas interfaces não é reconhecida como uma boa prática, que pode incorrer em diversos problemas.

De modo a contornar essa situação, para os diferentes ambientes de desenvolvimento e linguagens de programação, *drivers* tem sido desenvolvidos de modo a viabilizar a execução dos comandos no banco de dados. Em muitas situações, a decisão de qual combinação entre banco de dados não-relacional e *driver* a ser empregada pode ser um problema, devido a variedade de possibilidades, assim como o desconhecimento dos pontos positivos e negativos de cada solução.

Neste artigo, por meio de um estudo comparativo, busca-se avaliar as duas principais soluções de *drivers* para o MongoDB ¹, respectivamente MongoClient ² e Mongoose ³, em ambientes de aplicação Node.js. O principal fator considerado para a realização desse análise consiste na definição prévia de esquema para a manipulação dos dados em operações de CRUD (*create-read-update-delete*). Conceitualmente, os banco de dados não-relacionais não requerem esse predefinição, proporcionando flexibilidade, contudo não existe nada que impeça sua utilização, principalmente quanto a respeito do desempenho; possibilitando algum impacto relevante.

A escolha do banco de dados MongoDB ocorreu devido a sua enorme popularidade recente, empregado em diversas aplicações e linhas de pesquisa; o qual consiste na principal opção dentre os bancos de dados que adotam a estratégia de armazenamento orientada a documentos. A respeito do ambiente Node.js, apesar de uma tecnologia recentes alguns trabalhos apontam a sua viabilidade no desenvolvimento de aplicações [Chaniotis et al. 2015]; além disso, com essa escolha, tanto aplicação quanto banco de dados utilizam JavaScript, permitindo a elaboração de um sistema uniforme, em termos de linguagem de programação.

Na análise conduzida neste trabalho, o desempenho do banco de dados MongoDB atrelado a cada um dos *drivers* investigados é analisado sobre perspectivas de tempo de execução, tempo de uso de processamento e consumo de memória sob um conjunto de dados genérico, propiciando a identificação de algumas características.

O restante desse trabalho está organizado do seguinte modo: na Seção 2 é realiza uma revisão conceitual sobre banco de dados não-relacional, assim como sobre o MongoDB e seus drivers. Na Seção 3 é apresentado o ambiente de execução Node.js assim como detalhes sobre uso de memória que são relevantes no estudo comparativo. A Seção 4 é apresentado o estudo comparativo realizado neste trabalho. Seção 4.3 são apresentados os resultados quantitativos obtidos, cuja discussão é elaborada na Seção 5. Em seguida, na Seção 6 e, por fim, na Seção 8 as considerações finais desse trabalho.

2. Banco de Dados Não-Relacional

Os bancos de dados NoSQL, , foram desenvolvidos visando armazenar e processar grandes volumes de dados. Em linhas gerais os bancos de dados NoSQL são livres de

¹<https://www.mongodb.com/>

²<https://mongodb.github.io/node-mongodb-native/>

³<https://mongoosejs.com/>

esquematisações, lidam com dados não estruturados como e-mail, documentos e mídias sociais de maneira eficiente [Mohamed et al. 2014, Ramesh et al. 2016].

O termo NoSQL é comumente utilizado para se referir a uma ampla variedade de armazenamentos de dados nos quais as restrições de transação ACID foram relaxadas para permitir melhor dimensionamento e desempenho horizontal [Rafique et al. 2018]. Os recursos gerais presentes nos bancos de dados NoSQL são sumarizados em: esquemas menos estruturados, suporte a operações de junção, alta escalabilidade, modelagem de dados simples com linguagem de consulta simples [Ramesh et al. 2016]. Os bancos de dados NoSQL foram categorizados em: armazenamento de documentos, famílias de colunas, chave/valor, gráficos e multimodais [González-Aparicio et al. 2016].

Este trabalho tem como foco a categoria orientada a documentos, a qual permite a modelagem de dados estreitamente relacionados a programação orientada a objetos. Cada documento é considerado como um objeto, da mesma forma cada documento pode ser um JSON ou um XML no banco de dados orientado a documentos.

O conceito de esquema nos bancos de dados de documentos é dinâmico, uma vez que, cada documento pode conter campos distintos um dos outros, sendo útil na modelagem de dados não estruturados e polimórficos. Essa categoria permite consultas robustas, em que qualquer combinação de campos no documento pode ser realizada visando consultar dados [Patil et al. 2017]. Em termos de estruturação, seus dados são organizados em coleções de documentos, as quais utilizam uma estrutura semelhante a JSON (*JavaScript Object Notation*) ou XML (*Extensible Markup Language*).

2.1. MongoDB

O MongoDB [Membrey et al. 2011] é um banco de dados orientado a documentos de código-aberto. Embora não seja considerado um banco de dados relacional, esse apresenta muitas funcionalidades providas por essa categoria, como ordenação, indexação secundária e consultas de intervalo.

A estruturação dos dados não acontece com base em tabelas com suas respectivas colunas e tuplas, mas sim coleções de documentos. O banco de dados não impõe um esquema, contudo, normalmente todos os documentos em uma coleção são de propósito semelhante ou relacionado [Kanade et al. 2014, Lutu 2015]. Há duas abordagens para modelagem de documentos:

- Modelo de dados incorporado: os dados relacionados são incorporados em uma única estrutura ou documento. Esses esquemas são geralmente conhecidos como modelos “desnormalizados”.
- Modelo de dados normalizado: os dados possuem referências de documentos para registrar relacionamentos entre esses, mas a “união” de documentos deve ser feita diretamente no código-fonte da aplicação.

O armazenamento dos dados ocorre através da serialização de objetos Javascript, também conhecidos JSON, cuja implementação interna utiliza uma codificação binária chamada BSON [BSONSpecification 2019].

O banco de dados MongoDB disponibiliza diversos drivers para linguagens de programação como Java, C++, C#, PHP e Python [Lutu 2015], assim como para aplicações baseadas em Node.js.

Nesse contexto, dentre os drivers existentes, tem-se como destaque o MongoClient⁴, consiste na solução oficial e nativa provida organização, provendo um conjunto de funcionalidades que permite a manipulação dos dados e uso de recursos avançados do sistema. Essa solução é caracterizada pela modelagem documentos-objeto (*ODM – Object-Document Modeler*) de modo implícito ao banco de dados.

Assim como o anterior, o Mongoose consiste também em um driver para MongoDB, porém esse provê a modelagem de dados utilizando um mecanismo semelhante ao mapeamento de dados em tabelas (*ORM – Object Relational Mapping*) utilizado em banco de dados relacionais [Mardan 2014], executando diversas tarefas de verificação e validação dos dados, como nulidade ou tipagem, previamente definidos por meio da elaboração de um esquema.

3. Node.js

Node.js é uma plataforma construída sob o ambiente de execução para JavaScript do navegador Google Chrome, para criação facilitada de aplicações de internet rápidas e escaláveis [NodeJS 2019]. Esse ambiente se baseia em mecanismo orientado a eventos de entrada e saída não-bloqueante, o que viabiliza a interação do usuário enquanto demais tarefas executam em segundo-plano, resultando em aplicações leves e eficientes.

Atualmente, pode ser observado um crescente número de projetos utilizando essa tecnologia, o que é comprovado pela popularidade da plataforma desde as suas versões iniciais, atrelada a crescente utilização da linguagem Javascript⁵.

3.1. Organização de Memória em Aplicações Node.js

Toda aplicação Node.js em execução, assim como qualquer processo em geral, requer que uma região da memória seja reservada. A Figura 1 apresenta a organização da memória em processos Node.js.

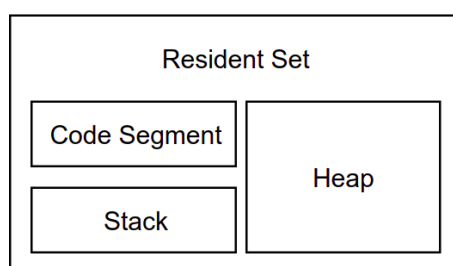


Figura 1. Organização de memória de um processo Node.js – adaptado de [Gagliardi 2018].

A primeira região, que engloba as demais, é chamada *Resident Set*, essa corresponde a toda memória utilizada no processo. Em seguida tem-se a região *Code Segment*, a qual armazenada todas as instruções definidas para o programa. A próxima região *Stack* armazena todas as variáveis e estruturas de dados utilizadas durante o tempo de vida dessas. Por fim, tem-se a região *Heap*, a qual armazena dados específicos como

⁴<https://mongodb.github.io/node-mongodb-native/index.html>

⁵<https://github.com/search?l=JavaScript&o=desc&q=stars%3A%3E1&s=stars&type=Repositories>

objetos, strings e closures; em geral, cada processo aloca esse região com um tamanho predefinido, contudo essa pode ser utilizada apenas parcialmente ou em sua totalidade [Gagliardi 2018].

4. Estudo Comparativo

Dentre as perspectivas de análise desse trabalho consiste na comparação de dois drivers, em que um utiliza definição previamente e outro não, assim

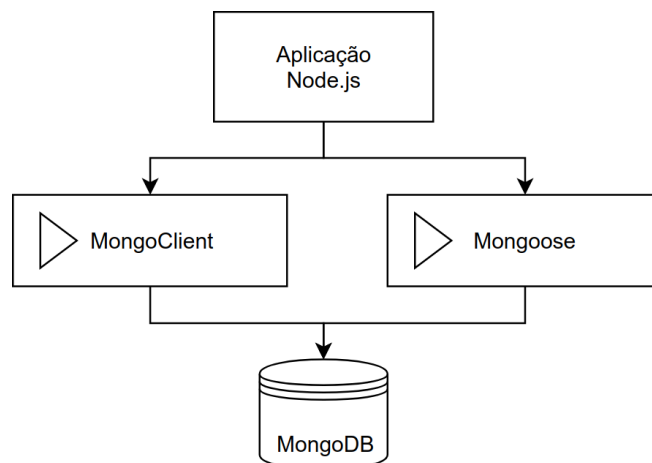


Figura 2. Estudo de caso para comparação dos drivers.

Q1 – O tamanho médio dos registros impacta de modo relevantes quanto ao uso de memória nas operações de CRUD?

Q2 – O tamanho médio dos registros pode influenciar no uso de CPU nas operações de CRUD?

Q3 – O tamanho médio dos registros impacta no tempo de execução de cada uma das operações de CRUD?

4.1. Dataset

O presente estudo utilizou um conjunto de dados com cerca 18 mil instâncias de dados, proveniente do seguinte dataset ⁶. Originalmente, todos os registros presentes são compostos por 89 atributos, predominante textuais, obtendo um tamanho médio de 1,37KB. A partir do conjunto original, foi construído um conjunto reduzido em número de atributos (6 atributos), com o mesmo total de instância, porém com tamanho médio de 0,13KB.

4.2. Ambiente de Execução

O ambiente de execução para os testes de desempenho consistiu em um computador pessoal com sistema operacional Ubuntu 18.04.2, processador Intel i3 3217U e memória RAM de 5GB.

Durante a execução dos testes, o ambiente de execução da aplicação Node.js foi definido o uso do *heap* de memória com limite máximo de 3GB, o que restringiu no limite superior do número de operações executadas em cada cenário de teste.

⁶<https://www.kaggle.com/karangadiya/fifa19>

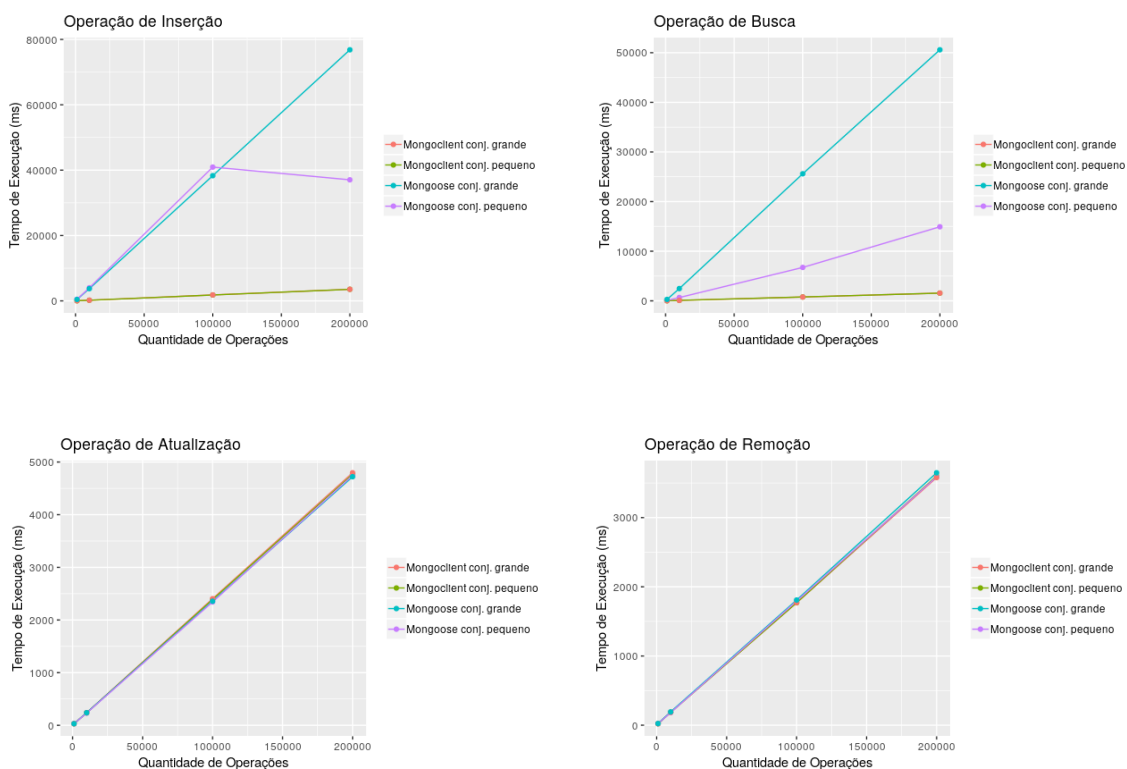


Figura 3. The average and standard deviation of critical parameters

Em cada cenário de execução, foram extraídos dados relativos ao tempo de execução, tempo de uso de CPU e uso de memória RAM. Foram analisados cenários com diferentes quantidades de operações CRUD realizadas, as quais variaram de 1000, 10000, 100000 e 200000; cada qual foi repetido 10 vezes.

A obtenção das métricas de desempenho de cada um dos cenários, para tempo de execução, tempo de uso de CPU e memória RAM, foi realizada através da biblioteca JSMeter ⁷.

4.3. Resultados

Esses são os resultados de tempo

Esses são os resultados de uso de cpu

Esses são os resultados de uso de memoria

⁷<https://github.com/wahengchang/js-meter>

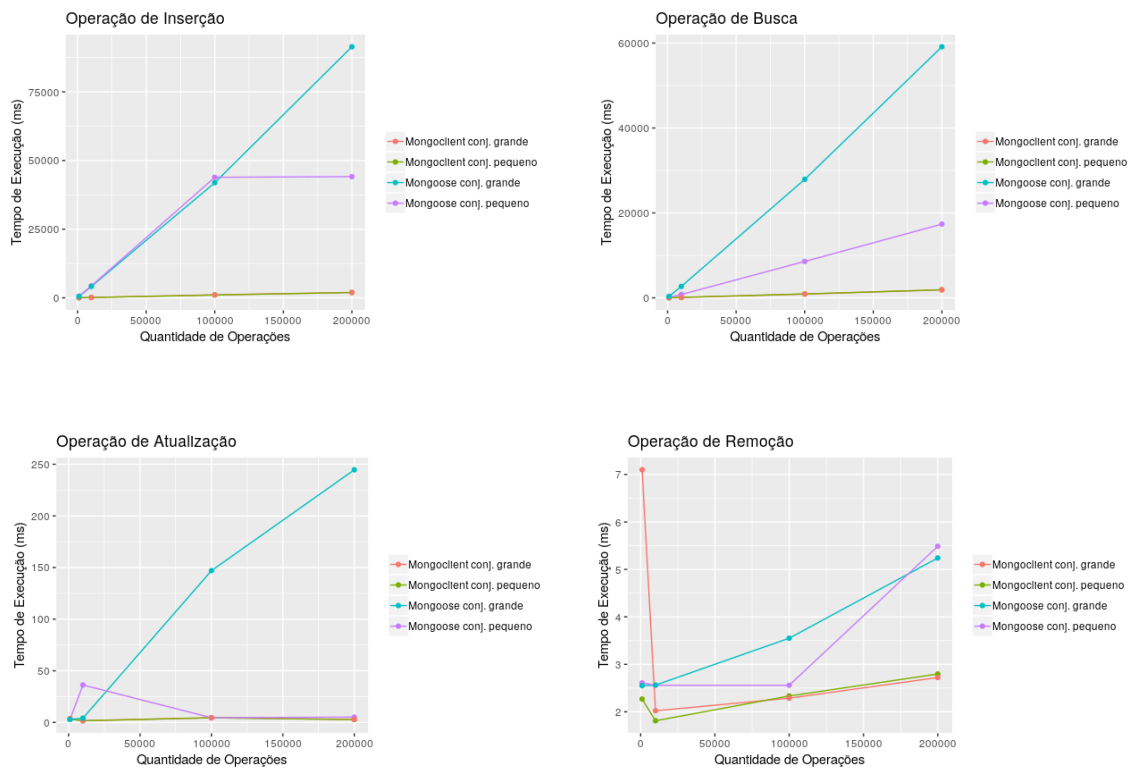


Figura 4. The average and standard deviation of critical parameters

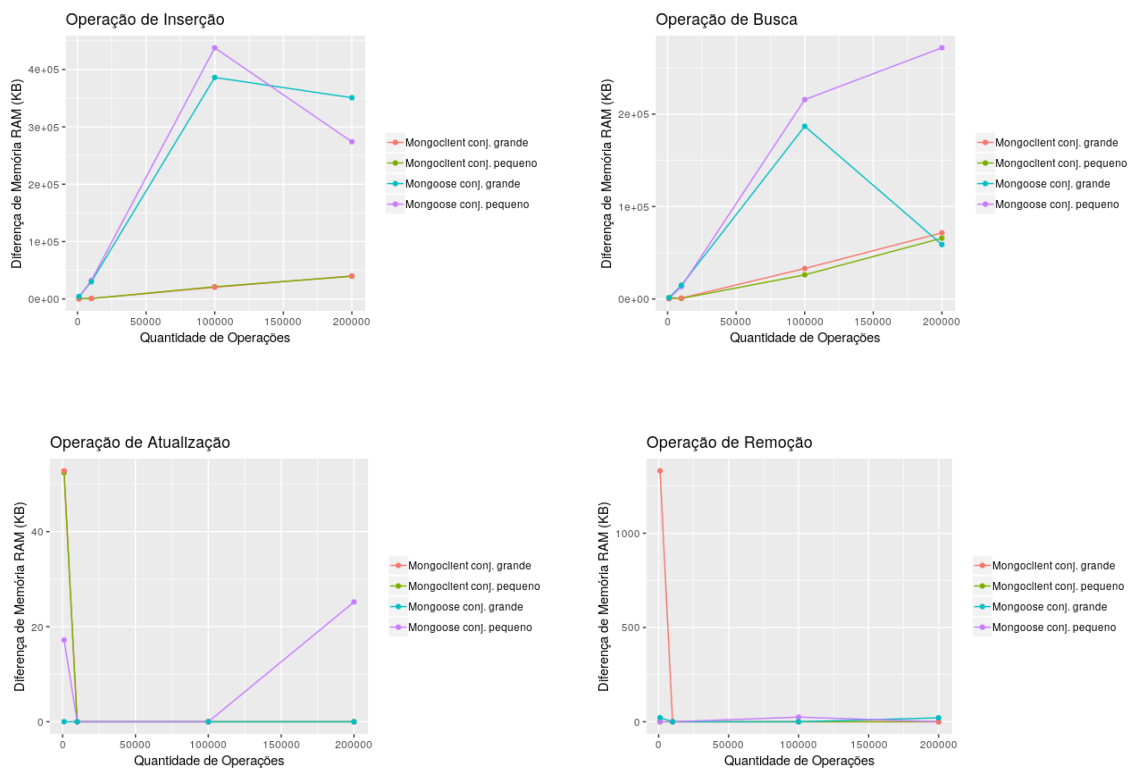


Figura 5. The average and standard deviation of critical parameters

5. Discussão dos Resultados

5.1. Discussão da Q1

5.2. Discussão da Q2

5.3. Discussão da Q3

6. Ameaças à Validação

7. Trabalhos Relacionados

[Kanade et al. 2014] did a study for NoSQL databases with both normalized and denormalized forms using a similar dataset, and have found that the embedded MongoDB data model provides a much better efficiency as compared to a normalized model.

8. Considerações Finais

9. Referências Bibliográficas

Referências

BSONSpecification (2019). Binary json.

Chaniotis, I. K., Kyriakou, K.-I. D., and Tselikas, N. D. (2015). Is node.js a viable option for building modern web applications? a performance evaluation study. *Computing*, 97(10):1023–1044.

Cooper, B. F., Silberstein, A., Tam, E., Ramakrishnan, R., and Sears, R. (2010). Benchmarking cloud serving systems with ycsb. In *Proceedings of the 1st ACM symposium on Cloud computing*, pages 143–154. ACM.

Edlich, S. (2015). Nosql-your ultimate guideto the non-relational universe! nosql.

Gagliardi, V. (2018). How to inspect the memory usage of a process in node.js.

González-Aparicio, M. T., Younas, M., Tuya, J., and Casado, R. (2016). A new model for testing crud operations in a nosql database. In *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, pages 79–86.

Han, J., Haihong, E., Le, G., and Du, J. (2011). Survey on nosql database. In *2011 6th international conference on pervasive computing and applications*, pages 363–366. IEEE.

Kanade, A., Gopal, A., and Kanade, S. (2014). A study of normalization and embedding in mongodb. In *2014 IEEE International Advance Computing Conference (IACC)*, pages 416–421. IEEE.

Lutu, P. (2015). Big data and nosql databases: new opportunities for database systems curricula. *Proceedings of the 44th Annual Southern African Computer Lecturers' Association, SACLA*, pages 204–209.

Mardan, A. (2014). Boosting your node.js data with the mongoose orm library. In *Practical Node.js*, pages 149–172. Springer.

Membrey, P., Plugge, E., and Hawkins, D. (2011). *The definitive guide to MongoDB: the noSQL database for cloud and desktop computing*. Apress.

Mohamed, M., G. Altrafi, O., and O. Ismail, M. (2014). Relational vs. nosql databases: A survey. *International Journal of Computer and Information Technology (IJCIT)*, 03:598.

NodeJS (2019). Node.js.

Patil, M. M., Hanni, A., Tejeshwar, C. H., and Patil, P. (2017). A qualitative analysis of the performance of mongodb vs mysql database based on insertion and retrieval operations using a web/android application to explore load balancing — sharding in mongodb and its advantages. In *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pages 325–330.

Rafique, A., Van Landuyt, D., Lagaisse, B., and Joosen, W. (2018). On the performance impact of data access middleware for nosql data stores a study of the trade-off between performance and migration cost. *IEEE Transactions on Cloud Computing*, 6(3):843–856.

Ramesh, D., Khosla, E., and Bhukya, S. N. (2016). Inclusion of e-commerce workflow with nosql dbms: Mongodb document store. In *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pages 1–5.

Ward, J. S. and Barker, A. (2013). Undefined by data: a survey of big data definitions. *arXiv preprint arXiv:1309.5821*.