Clusters Big Data utilizando *Raspberry Pi* e *Apache Hadoop* - Uma Quasi-Revisão Sistemática da Literatura

1st Antônio José A. Neto*, 2nd José M. dos Santos*, 3rd José A. C. Neto§, 4rd Edward D. Moreno*
*Departamento de Computação - Universidade Federal de Sergipe (UFS) - São Cristóvão - Sergipe
§Coordenação de Computação - Instituto Federal de Sergipe (IFS) - Itabaiana - Sergipe
{antonio.neto, edward}@dcomp.ufs.br, marcelotos@academico.ufs.br, jose.neto@ifs.edu.br

***Resumo*—Este trabalho tem como objetivo identificar como estão sendo desenvolvidos os *clusters big data* de baixo custo, utilizando *Raspberry Pi* e *Apache Hadoop*, e como os mesmos estão sendo validados e monitorados. Para tal fim, foi elaborada uma Quasi-Revisão Sistemática da Literatura (QRSL), resultando em 9 artigos relevantes aptos a responder 3 questões de pesquisa. A QRSL identificou que os modelos de *Raspberry Pis* mais utilizados no desenvolvimento dos *clusters* são a *Raspberry Pi 4B* e a *Raspberry Pi 2B*, e que para sua validação os *benchmarks Terasort* e *Wordcount* são os mais citados na literatura, seguidos da abordagem original do *Map Reduce* e o *TestDFSIO*. As 3 únicas ferramentas encontradas para monitoramento dos recursos do *cluster* foram a *Ganglia*, *Grafana* e a *Prometheus*.**

***Index Terms*—*Cluster*, *Big Data*, *Raspberry Pi*, *Apache Hadoop*, *Benchmarks*, Revisão Sistemática**

I. Introdução

O termo *big data* é definido como os ativos de informações de alto volume, alta velocidade e/ou alta variedade, que exigem formas inovadoras e eficazes de processamento de informações, permitindo *insights* aprimorados, tomada de decisões e automação de processos [1].

Com os avanços na utilização de *big data* e *data science*, são exigidos, cada vez mais, recursos computacionais para processar esses grandes volumes de dados, tornando-se indispensável o uso de computação de alto desempenho (*High Performance Computing* - HPC).

A HPC refere-se a um sistema de computação que inclui vários processadores como parte de uma única máquina ou um *cluster* trabalhando como um recurso individual. A HPC deve sua característica de computação de alta velocidade à sua grande capacidade de processar informações [2]. Em contra partida, por conta do seu alto custo, tem se buscado alternativas de soluções mais baratas e eficazes. Atualmente, com os avanços de *hardware* que o mundo está presenciando, a *Raspberry Pi* tem trazido oportunidades de implantações de *clusters* econômicos e energeticamente eficientes.

A *Raspberry Pi* é um Computador de Placa Única (*Single Board Computer* - SBC) desenvolvido para promover a ciência da computação na educação [3]. Os *clusters* formados por esses dispositivos, atrelado ao uso do *Apache Hadoop*, têm se mostrado uma solução viável e econômica para a realização de tarefas que envolvem o uso de *big data*.

A montagem de um *cluster* formado por esse tipo de equipamento (Figura 1) não requer grandes investimentos para sua implantação. Além disso, a *Raspberry Pi* permite o uso de diversas bibliotecas de paralelismo, bem como o desenvolvimento de diversos algoritmos paralelos utilizando linguagens de programação [4] e os pesquisadores vêm convergindo para a ideia de que a melhor forma de implementar um *cluster* de *big data* de baixo custo é combinando *Raspberry Pi* com *Apache Hadoop* [5]–[8].

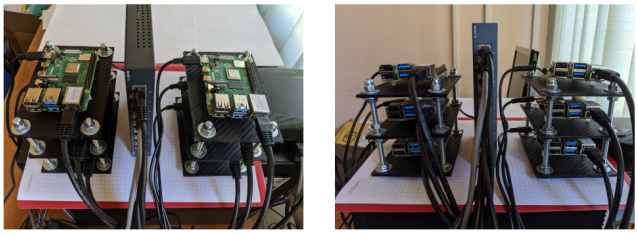

Figura 1. *Cluster Raspberry Pi*. Adaptado de Komninos *et al*. [9].

O *Apache Hadoop* é uma estrutura que permite o processamento distribuído de grandes conjuntos de dados em *clusters*, utilizando modelos de programação simples. Ele foi projetado para escalar desde servidores únicos até milhares de máquinas, cada uma, oferecendo computação e armazenamento locais. Ao invés de depender de recursos de *hardware* para fornecer alta disponibilidade, a própria biblioteca foi projetada para detectar e lidar com falhas na camada de aplicação, oferecendo um serviço altamente disponível em um *cluster* [10].

Seguindo essa vertente, este trabalho tem como objetivo identificar como estão sendo desenvolvidos os *clusters big data* de baixo custo, utilizando *Raspberry Pi* e *Apache Hadoop*, e como os mesmos estão sendo validados e monitorados. Para tal fim, foi elaborada uma Quasi-Revisão Sistemática da Literatura (QRSL), expondo os estudos relevantes nessa área de pesquisa.

O restante deste o trabalho está organizado da seguinte forma: A Seção II apresenta a metodologia utilizada. O estado da arte é descrito na Seção III. Os resultados encontrados através do método de pesquisa, filtragem e análise dos artigos selecionados são mostrados na Seção IV. A Seção V descreve as ameaças à validade deste trabalho. As considerações finais

sobre os resultados encontrados, bem como os trabalhos futuros sugeridos são apresentados na Seção VI, e por fim, os agradecimentos são feitos na Seção VII.

II. Metodologia

Uma QRSL é um estudo secundário baseado em evidências, que tem como proposta, o levantamento, a identificação, a avaliação e a classificação da literatura para responder questões de pesquisas específicas [11]. Diante disso, o protocolo utilizado neste trabalho segue as diretrizes de revisão sintetizadas por Kitchenham [11]. Embora a QRSL cumpra todas as fases de avaliação da revisão sistemática, ela não é considerada como tal pela falta de uma linha base para meta-análise, lhe faltando em seu estudo, modelos de comparação para os resultados [12].

Esta QRSL foi desenvolvida em três fases. São elas: Planejamento, Condução e Relatório Final da Revisão (Figura 2).

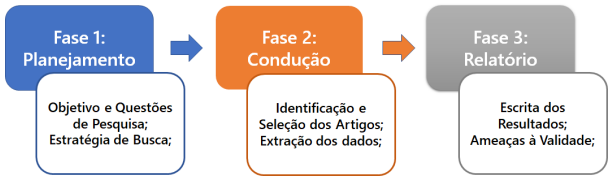

Figura 2. Fases do Planejamento da QRSL.

A. Planejamento

O planejamento foi iniciado com a definição do objetivo e das questões de pesquisa. Em seguida, a estratégia de busca foi construída, determinando as bases, os termos de busca e os critérios de inclusão e exclusão para a seleção dos artigos.

1) Objetivo e Questões de Pesquisa (QP): Este estudo tem o objetivo de levantar e avaliar como estão sendo desenvolvidos os *clusters big data* de baixo custo, utilizando *Raspberry Pi* e *Apache Hadoop*, e como os mesmos estão sendo validados e monitorados. Para alcançar esse objetivo foram definidas as seguintes questões de pesquisa:

- QP1: Quais são os modelos de *Raspberry Pis* que são utilizados para o desenvolvimento do *cluster big data*?
- QP2: Quais são os *benchmarks*/algoritmos/técnicas que estão sendo utilizados no *cluster big data*?
- QP3: Quais ferramentas são utilizadas para monitorar os recursos do *cluster*?

2) Bases de Dados: As bases de dados utilizadas neste estudo foram a *Scopus* e a *Web of Science*. As mesmas foram utilizadas pelo fato de indexarem as principais bases de dados científicas, além de seus motores de busca automatizados possuírem uma maior eficiência em encontrar estudos relevantes [11]. É importante destacar que todas as bases de dados foram acessadas através do Portal de Periódicos da CAPES [13].

3) Termos de Busca: Após a definição das bases de dados para encontrar os estudos primários, foi estabelecida a *string* de busca, com finalidade de identificação e classificação desses estudos, conforme pode ser visto na Figura 3. Nas duas bases de dados utilizadas foram empregados recursos de busca avançada, com o propósito de encontrar as palavras-chave pesquisadas, contidas nos títulos e/ou nos resumos do material prospectado.

```
(big data) AND (cluster) AND
(raspberry) AND (hadoop)
```

Figura 3. *String* de Busca.

4) Critérios de Inclusão e Exclusão: Os critérios de inclusão e exclusão são utilizados para delimitar estudos mais relevantes que respondam, ou não, as questões de pesquisa. Foram definidos os seguintes critérios de inclusão (CI):

- CI1: Publicações focadas no desenvolvimento de um *cluster big data* usando *Raspberry Pi* e *Apache Hadoop*;
- CI2: Publicações que contribuam para responder pelo menos uma das QPs;
- CI3: Trabalhos publicados entre os anos de 2015 e 2022;
- CI4: Trabalhos escritos nos idiomas Português ou Inglês;

E os critérios de exclusão (CE) foram:

- CE1: Publicações inconsistentes com o tópico de pesquisa;
- CE2: Publicações duplicadas;
- CE3: Publicações indisponíveis para *download*;

B. Condução da Revisão

A busca dos trabalhos foi executada no final do primeiro semestre de 2022. A quantidade de artigos identificados, o método de filtragem e a extração dos dados são detalhados a seguir.

A busca retornou 17 artigos, que em uma primeira análise, foram verificados o título e o resumo desses estudos, atentando-se ao objetivo do trabalho e à sua conclusão, identificando o alinhamento do artigo ao escopo da QRSL, resultando em 10 artigos. Posteriormente, os 10 artigos foram analisados por completo, avaliando a capacidade dos mesmos em responder ao objetivo dessa QRSL. Dentre todos os trabalhos, apenas 1 não atendeu aos requisitos [14], sendo removido pelo fato de que o estudo diverge do escopo desta QRSL. Todo esse processo metodológico desta pesquisa pode ser visualizado através da Figura 4.

C. Relatório da Revisão

Ao final de todo o processo metodológico da QRSL, foram considerados 9 artigos como relevantes. Uma síntese de cada um dos trabalhos selecionados é apresentada na Seção III.

Os artigos analisados por esta QRSL, demonstraram o quão importante tem sido essa discussão e como a temática desta pesquisa tem sido explorada e pesquisada no mundo.

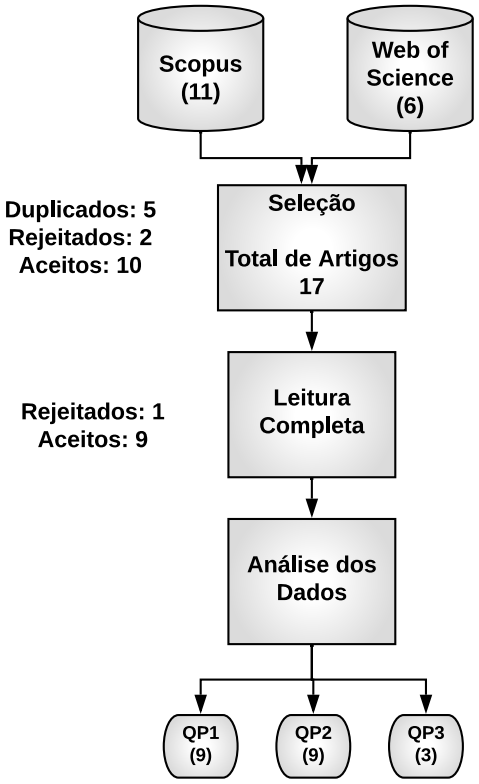

Figura 4. Processo de Seleção dos Artigos.

A Figura 5 demonstra a distribuição das participações nas publicações por continente, sendo a Europa o continente com maior percentual de publicações dentre os demais. Alguns trabalhos apresentam uma multinacionalidade entre os autores, resultando em mais de um país representado nos trabalhos encontrados.

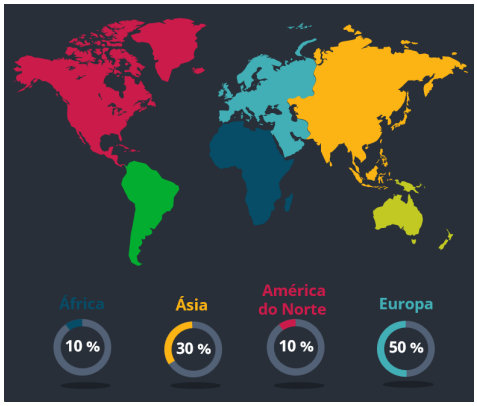

Figura 5. Distribuição das publicações por continente.

III. Estado da Arte

A. Big Data Processing on Single Board Computer Clusters: Exploring Challenges and Possibilities - Lee et al. [6]

Nesse trabalho, os autores exploraram os desafios das últimas gerações de *Raspberry Pis* utilizados em *clusters big data*. Eles construíram 1 *cluster big data* composto por 6 nós (5 escravos e 1 mestre), onde em cada nó escravo utiliza 1 *Raspberry Pi 4B* e o nó mestre utiliza 1 *Desktop* (Processador i5 de 3.7GHz (6 *cores*), *Solid State Drive* (SSD) de 500 *Gigabytes* (GB) e Memória RAM de 8 GB) e *Apache Hadoop*.

Os autores realizaram uma bateria de testes, a fim de avaliar a performance de uma única *Raspberry Pi*, assim como a performance do *cluster* desenvolvido, utilizando alguns dos principais *benchmarks*: *Wordcount*, *Terasort*, *TestDFSIO* e o *Pi Quasi-Monte Carlo*.

Eles comparam o *cluster* desenvolvido com o *Desktop* e com 1 *cluster big data* composto por *Raspberry Pi 3B* (1 nó mestre e 5 nós escravos), além de estudarem o impacto da performance no armazenamento de dados utilizando 3 diferentes tipos de dispositivos: *Micro Storage Device Card* (MicroSD *Card*) de 32 GB; MicroSD *Card* 64 GB; *Universal Flash Storage* (UFS) de 256 GB.

Os mesmos concluíram que ao utilizar uma versão mais recente do *hardware Raspberry Pi*, o *cluster* tem uma melhor capacidade de processamento, bem como a utilização do dispositivo UFS, como unidade de armazenamento, tem um melhor *ratio* (*Input/Output*) em relação ao uso de MicroSD *Card*.

B. Hadoop Performance Analysis on Raspberry Pi for DNA Sequence Alignment - Turana et al. [15]

Nessa pesquisa, os autores buscaram analisar a sequência de alinhamento de DNA. Para isso, foi construído 1 *cluster Hadoop* na *Raspberry Pi B* usando o hardware da *Raspberry* como *commodity*. Foram utilizados 6 nós no *cluster* (1 nó mestre e 5 nós escravos). Eles usaram a biblioteca nativa *Biodoop* e o *benchmark Wordcount* para os testes, além de utilizar a ferramenta *Ganglia* para o monitoramento dos recursos do *cluster*.

Os autores realizaram uma comparação entre o *cluster Raspberry Pi B* e 6 *Desktops* (processador *dual core* de 1,86 GHz, *Hard Disk Drive (HDD)* de 160 GB e Memória RAM de 1 GB).

Foi concluído que é possível utilizar o *hardware* alternativo de baixo custo para implementar o *cluster Hadoop*, mesmo com a desvantagem do aumento de tempo na conclusão dos experimentos comparados.

C. Understanding the Performance of Low Power Raspberry Pi Cloud for Big Data - Haiji & Tso [5]

Esse trabalho apresenta um conjunto de experimentos para testar a performance de 1 nó único e de 1 *cluster* composto por 12 *Raspberry Pis 2B* (1 mestre e 11 escravos) com e sem ambiente de virtualização, utilizando a ferramenta *Docker* para estudar a viabilidade do mesmo para *big data analyics* em tempo real.

Utilizando os *benchmarks Wordcount* e *Terasort*, os autores executaram uma bateria de testes com diversas configurações de tamanhos de arquivos, variando entre 1 GB a 6 GB. Os mesmos avaliaram que em um ambiente virtualizado, o consumo de CPU e memória RAM torna-se maior, a taxa de

transferência da rede diminui e os picos de acessos ocorrem com menos frequência e menos intensidade.

D. Towards Green Data Centers - Bourhnane et al. [16]

No artigo mencionado, foi implementado um comparativo de *hardware* no uso do *Apache Hadoop*. Enquanto um experimento usou 1 *Desktop* (2 processadores *Core 2 Duo de 2.20 GHz, HDD de 160GB*), no outro foi utilizado 1 *cluster* com 5 *Raspberry Pis 3B+*, sendo 1 nó mestre e 4 nós escravos. Esse *cluster* tinha como diferencial, o uso de 5 HDDs de 1 TB (externos), além de 5 MicroSD *Card* (internos) de 8 GB.

A finalidade da pesquisa era agregar uma motivação no uso da prática de computação eficiente e ecológica, ou seja, na computação verde. Para tanto, foram executados testes com os *benchmarks TestDFSIO* e *Terasort* para a análise de desempenho e eficiência energética do *cluster*, utilizando a *Raspberry Pi*.

Os experimentos mostraram desempenho significativo com o *TestDFSIO* em relação ao *cluster* tradicional. No entanto, o *Terasort* forneceu um menor desempenho, que pode ser facilmente superado adicionando mais nós ao *cluster*. Nos testes de consumo energético no *cluster Raspberry Pi*, foi comprovado o baixo consumo de energia, provando ser esse, um experimento de computação verde.

E. An Efficient Implementation of Mobile Raspberry Pi Hadoop Clusters for Robust and Augmented Computing Performance - Srinivasan et al. [17]

Nesse trabalho, os autores utilizaram 6 *clusters* diferentes, sendo eles configurados com 5 a 10 *Raspberry Pis 3B*, sempre utilizando o padrão de 1 nó mestre e os nós remanescentes como escravos. Também foi utilizado 1 *Desktop* (processador i5, 3.1 GHz, Memória RAM de 8 GB) para os mesmos experimentos.

Foi explorado pelos autores, a implementação do *Hadoop*, tanto para o *cluster* construído com *Raspberry Pi* quanto para o *Desktop*. Tais experimentos visavam testes de extração de pontos de recurso em uma imagem utilizando o algoritmo *Surf* e comparando os resultados obtidos.

Foi verificado pelos pesquisadores que os *clusters* formados por *Raspberry Pi* têm melhor performance que o *Desktop* para grandes *datasets*.

F. A containerized big data streaming architecture for edge cloud computing on clustered single-board devices — A Containerized Edge Cloud Architecture for Data Stream Processing - Scolati et al. [18], [19]

Em ambos artigos foram realizados experimentos em 1 *cluster* de 8 *Raspberry Pis 2B* com o gerenciador *Docker Swarm*. Ao *cluster* foi conferida a seguinte configuração: 1 nó mestre, 4 nós escravos e 3 nós para coleta de dados do *cluster*.

Os autores implementaram o *cluster* utilizando o *Docker* para hospedar o *Apache Hadoop* e o *Apache Spark* como *containers* e, dessa forma, aprimorar o processamento de *streaming de dados*. Para monitorar os recursos do *cluster*, os mesmos usaram as ferramentas *Prometheus* e *Grafana*.

Eles avaliaram o desempenho dos mesmos, mostrando que o uso da "conteinerização" aumenta a tolerância à falhas e a facilidade de manutenção.

G. Scale-down Experiments on TPCx-HS - Böther & Rabl [7]

Nesse estudo, os autores fizeram a execução do *benchmark* TPCx-HS em 2 *clusters* utilizando *Raspberry Pi*: O primeiro *cluster* composto por 4 *Raspberry Pis 3B+* trabalhando como escravos e 1 *Raspberry Pi 4B* como mestre, utilizando *Micro SD Cards* de 32 GB em cada dispositivo e o segundo *cluster* composto por 5 *Raspberry Pis 4B* (1 mestre e 4 escravos), onde cada *Raspberry* utilizou MicroSD *Card* de 128 GB para armazenamento.

Nos experimentos, foram utilizadas 2 configurações diferentes nos tamanhos dos arquivos usados pelo *benchmark* TPCx-HS: 1 GB e 10 GB. Em seguida, foi feita a comparação dos resultados obtidos com os disponíveis no site[1] do *benchmark*.

Os autores concluíram que a geração de *Raspberry Pi 4B* resolve o gargalo de memória RAM que existia na geração de *Raspberry Pi 3B+*. Além disso, os mesmos defendem o uso de *clusters Raspberry Pi* devido ao seu custo-benefício.

H. Performance of Raspberry Pi microclusters for Edge Machine Learning in Tourism - Komninos et al. [9]

No desenvolvimento dessa pesquisa foram utilizadas 6 *Raspberry Pis 4B*, com 4 GB de RAM e cartão MicroSD *Card* de 64 GB em cada nó. No *cluster* foram instalados o *Hadoop* para armazenamento distribuído de arquivos e o *Spark* para aprendizado de máquina.

Tal estudo tinha como objetivo utilizar recursos computacionais em uma arquitetura distribuída de baixo custo, para atender aplicações turísticas através da análise de *big data*.

Após experimentos e testes, foi demonstrado que o desempenho do *cluster* foi suficiente para fins de treinamento e execução de modelos de aprendizado de máquina em um contexto de computação de borda.

IV. Resultados

Nesta seção, os resultados da extração de dados são expostos, analisados e discutidos e, para um melhor entendimento e visualização dos mesmos, foi feita uma subdivisão de acordo com as QPs.

A. QP1: Quais são os modelos de Raspberry Pis que são utilizados para o desenvolvimento do cluster?

A QRSL apontou 5 diferentes modelos de *Raspberry Pis* utilizados para o desenvolvimento dos *clusters* citados nos trabalhos encontrados, sendo que o *Raspberry Pi 2B* e *Raspberry Pi 4B* foram os modelos mais utilizados nos experimentos. Cada um foi citado em 3 artigos. Alguns trabalhos utilizaram mais de 1 modelo de *Raspberry Pi*, geralmente para fazer algum comparativo entre os resultados encontrados.

É possível observar que quanto mais recente é o trabalho, mais recente é a versão da *Raspberry Pi* (com exceção dos trabalhos de Scolati *et al*. [18], [19]), justificando-se pelo fato

[1] https://www.tpc.org/tpcx-hs/results/tpcxhs_perf_results5.asp?version=2

de que quanto mais nova é a versão do dispositivo, mais poderoso computacionalmente ele é.

Todos os modelos encontrados podem ser visualizados na Tabela I. Um comparativo entre os modelos existentes de *Raspberry Pis* pode ser encontrado no trabalho de SocialCompare [20].

Tabela I
MODELOS DE RASPBERRY PIS UTILIZADOS NO DESENVOLVIMENTO DO CLUSTER

Modelo	Referência
Raspberry Pi 4B	[6], [7], [9]
Raspberry Pi 3B+	[7], [16]
Raspberry Pi 3B	[6], [17]
Raspberry Pi 2B	[5], [18], [19]
Raspberry Pi B	[15]

B. QP2: Quais são os benchmarks/algoritmos/técnicas que estão sendo utilizados no cluster big data?

Foram identificados 9 *benchmarks*/algoritmos/técnicas nos artigos selecionados, sendo o *Terasort* e o *Wordcount* os 2 mais utilizados, com 3 citações cada. Alguns estudos mencionaram a utilização de mais de 1 *benchmark*/algoritmo/técnica.

A incidência da utilização dos algoritmos/*benchmarks Terasort* e o *Wordcount*, juntamente com o *TestDFSIO* e o *Pi Quasi-Monte Carlo*, deve-se ao fato de que os mesmos estão disponíveis nas distribuições do *Apache Hadoop*, sendo tecnicamente mais cômodo aos pesquisadores utilizarem *benchmarks*/algoritmos/técnicas padrões (*default*), do que a instalação de um novo.

É importante salientar que o algoritmo *Map-Reduce* citado nos trabalhos de Scolati *et al.* [18], [19] é baseado em sua abordagem original, diferentemente da abordagem do *Map-Reduce* do *Apache Hadoop*.

Todos os *benchmarks*/algoritmos/técnicas mencionados nos trabalhos encontrados podem ser observados na Tabela II.

Tabela II
BENCHMARKS/ALGORITMOS/TÉCNICAS UTILIZADAS NOS CLUSTERS BIG DATA

Benchmark/Algoritmo/Técnica	Referência
Árvore de Decisão	[9]
Map-Reduce	[18], [19]
Pi Quasi-Monte Carlo	[6]
Regressão Linear	[9]
SURF	[17]
Terasort	[5], [6], [16]
TestDFSIO	[6], [16]
TPCx-HS	[7]
Wordcount	[5], [6], [15]

C. QP3: Quais ferramentas são utilizadas para monitorar os recursos do cluster?

Para resposta da QP3, apenas 3 trabalhos mencionaram o uso de alguma ferramenta para monitorar os recursos do *cluster*. Foram elas: *Ganglia*[2], *Grafana*[3] e *Prometheus*[4], conforme pode ser observado na Tabela III.

Todos os *softwares* são de código aberto (*open source*). O *Ganglia* é um sistema de monitoramento distribuído e escalável para uso na computação de alto desempenho, como *clusters* e computação em *grids*. O *Grafana* é uma plataforma para visualizar e analisar métricas por meio de *dashboards* personalizados. Ele tem suporte para diversos tipos de bancos de dados e pode ser instalado em qualquer sistema operacional. O *Prometheus* é uma solução de monitoramento para gravar e processar qualquer série temporal puramente numérica. Ele reúne, organiza e armazena métricas juntamente com identificadores exclusivos de data/hora.

Em alguns trabalhos, os autores mencionaram o fato de monitorar os recursos do *cluster* sem utilizar uma ferramenta específica para tal [5]–[7], [16].

Tabela III
FERRAMENTAS UTILIZADAS PARA O MONITORAMENTO DOS RECURSOS DO CLUSTER

Ferramenta	Referência
Ganglia	[15]
Grafana	[18], [19]
Prometheus	[18], [19]

V. AMEAÇAS À VALIDADE

Para que um trabalho seja aceito como contribuição ao conhecimento científico, o pesquisador precisa convencer os leitores de que as conclusões tiradas de um estudo empírico são válidas [21]. Nesta seção estão expostas todas as ameaças à validade encontradas neste estudo.

A. Validade de Construção

A *string* de busca e as questões de pesquisa podem não cobrir todos os estudos relevantes da área. Como forma de mitigação, as palavras-chave e as questões de pesquisa foram elaboradas a partir de uma leitura prévia do assunto, através do trabalho de Neto *et al.* [8].

B. Validade Interna

Caso seja executada novamente, esta QRSL poderia apresentar um resultado diferente do obtido. Para mitigar essa ameaça, a metodologia definida por Kitchenham [11] foi usada nesse trabalho. Conflitos na seleção de artigos foram discutidos entre os autores, para atenuar o viés pessoal na seleção dos estudos.

C. Validade Externa

O resultado da busca pode não conter todos os artigos relevantes para o estudo. Como forma de diminuir essa ameaça, a pesquisa foi realizada em duas das principais bases de dados científicas, *Web of Science* e *Scopus*.

[2]http://ganglia.sourceforge.net/
[3]https://grafana.com/
[4]https://prometheus.io/

VI. Considerações Finais

Neste trabalho, uma QRSL foi elaborada com o objetivo de levantar e avaliar como estão sendo desenvolvidos os *clusters big data* de baixo custo, utilizando *Raspberry Pi* e *Apache Hadoop*, e como os mesmos estão sendo validados e monitorados, através de 3 questões de pesquisa. Para responder essas questões, foram feitas buscas nas bases de dados *Scopus* e *Web of Science*, que retornaram 17 trabalhos. Após a aplicação do protocolo da QRSL, apenas 9 estudos foram classificados como relevantes.

Foram encontrados 5 modelos diferentes de *Raspberry Pis*, sendo o *Raspberry Pi 2B* e *Raspberry Pi 4B* os modelos mais utilizados nos experimentos e cada um citado em 3 artigos (QP1).

A respeito da QP2, foram identificados 9 *benchmarks*/algoritmos/técnicas nos artigos selecionados, sendo o *Terasort* e o *Wordcount* os 2 mais utilizados, com 3 citações cada. Alguns trabalhos mencionam o uso de mais de 1 *benchmark*/algoritmo/técnica.

Por fim, como resposta para a QP3, foram encontradas 3 ferramentas para monitoramento dos recursos do *cluster* (*Ganglia*, *Grafana* e *Prometheus*) em 3 dos 9 artigos. Alguns autores mencionaram o fato de monitorar os recursos do *cluster* sem a utilização de uma ferramenta específica.

Essa QRSL mostrou um leque significativo de experimentos no uso dos *clusters big data* de baixo custo, utilizando *Raspberry Pi* e *Apache Hadoop*. Além desta pesquisa mostrar resultados relevantes, fornecendo assim subsídios para um melhor entendimento sobre o desenvolvimento desses *clusters*, este artigo pode auxiliar outros pesquisadores para o desenvolvimento de novos estudos nessa área de pesquisa.

Como trabalho futuro, está sendo desenvolvido 1 *cluster big data* de baixo custo composto por 9 *Raspberry Pis 4B* (1 mestre e 8 escravos) e *Apache Hadoop*. Além disso, será utilizado 1 *Raspberry Pi 3B+* como servidor de monitoramento dos recursos do *cluster*, utilizando as ferramentas *Zabbix*[5] e *Grafana*, devido ao fato de que as mesmas são ferramentas *open-source* e os autores possuírem um conhecimento prévio na utilização das mesmas.

A priori, será utilizado nos experimentos o armazenamento (MicroSD *Card*) de 16 GB e posteriormente o armazenamento de 128 GB. Serão executados alguns dos *benchmarks* encontrados nesta QRSL, comparando os resultados obtidos com os resultados apresentados pelos *clusters* citados nesta QRSL, além de avaliar o comportamento dos recursos (CPU, RAM, Temperatura, etc.) do *cluster* desenvolvido.

VII. Agradecimentos

Agradecemos à todos os revisores que se dispuseram a avaliar este trabalho, melhorando a qualidade desta pesquisa.

Referências

[1] Gartner, "Big data," Disponível em: https://www.gartner.com/en/information-technology/glossary/big-data. Acesso em: 18 de maio 2022, 2022.

[2] A. Middleton and P. Solutions, "Hpcc systems: Introduction to hpcc (high-performance computing cluster)," *White paper, LexisNexis Risk Solutions*, 2011.

[3] P. Giger, S. Srikugan, and B. L. Persaud, "A Raspberry Pi Cluster for Teaching Big-Data Analytics," Master's thesis, Universität Zürich, 2020.

[4] N. M. Mwasaga and M. Joy, "Implementing micro high performance computing (μhpc) artifact: Affordable hpc facilities for academia," in *2020 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2020, pp. 1–9.

[5] W. Hajji and F. P. Tso, "Understanding the performance of low power raspberry pi cloud for big data," *Electronics (Switzerland)*, vol. 5, no. 2, 2016. [Online]. Available: https://doi.org/10.3390/electronics5020029

[6] E. Lee, H. Oh, and D. Park, "Big Data Processing on Single Board Computer Clusters: Exploring Challenges and Possibilities," *IEEE Access*, vol. 9, pp. 142 551–142 565, 2021.

[7] M. Böther and T. Rabl, "Scale-down experiments on tpcx-hs," in *Proceedings of the International Workshop on Big Data in Emergent Distributed Environments*, ser. BiDEDE '21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3460866.3461774

[8] A. J. A. Neto, A. C. Neto, and E. D. Ordonez, "Low-cost clusters on big data - a systematic study," in *Proceedings of the Euro American Conference on Telematics and Information Systems*, ser. EATIS '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3544538.3544635

[9] A. Komninos, I. Simou, N. Gkorgkolis, and J. D. Garofalakis, "Performance of raspberry pi microclusters for edge machine learning in tourism," in *AmI*, 2019.

[10] A. S. Foundation, "Apache hadoop," https://hadoop.apache.org, 2022.

[11] B. Kitchenham, "Procedures for performing systematic reviews," *Keele University Technical Report TR/SE-0401*, vol. 33, 08 2004.

[12] G. H. Travassos, P. S. M. dos Santos, P. G. Mian, P. G. M. Neto, and J. Biolchini, "An environment to support large scale experimentation in software engineering," in *13th IEEE International Conference on Engineering of Complex Computer Systems (iceccs 2008)*, 2008, pp. 193–202.

[13] CAPES/MEC, "Portal de periódicos da capes," http://www.periodicos.capes.gov.br/, 2022.

[14] J. Lin, "Scaling down distributed infrastructure on wimpy machines for personal web archiving," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15 Companion. New York, NY, USA: Association for Computing Machinery, 2015, p. 1351–1355. [Online]. Available: https://doi.org/10.1145/2740908.2741695

[15] J. S. Turana, H. Sukoco, and W. A. Kusuma, "Hadoop performance analysis on raspberry pi for dna sequence alignment," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 14, no. 3, pp. 1059–1066, 2016.

[16] S. Bourhnane, M. R. Abid, R. Lghoul, K. Zine-Dine, N. Elkamoun, and D. Benhaddou, "Towards green data centers," in *Sustainable Energy for Smart Cities*, J. L. Afonso, V. Monteiro, and J. G. Pinto, Eds. Cham: Springer International Publishing, 2020, pp. 291–307.

[17] K. Srinivasan, C. Y. Chang, C. H. Huang, M. H. Chang, A. Sharma, and A. Ankur, "An efficient implementation of mobile Raspberry Pi Hadoop clusters for Robust and Augmented computing performance," *Journal of Information Processing Systems*, vol. 14, no. 4, pp. 989–1009, 2018.

[18] R. Scolati, I. Fronza, N. El Ioini, A. Samir, and C. Pahl, "A containerized big data streaming architecture for edge cloud computing on clustered single-board devices," *CLOSER 2019 - Proceedings of the 9th International Conference on Cloud Computing and Services Science*, no. May, pp. 68–80, 2019.

[19] R. Scolati, I. Fronza, N. El Ioini, A. Samir, H. R. Barzegar, and C. Pahl, "A Containerized Edge Cloud Architecture for Data Stream Processing," *Communications in Computer and Information Science*, vol. 1218 CCIS, no. May, pp. 150–176, 2020.

[20] SocialCompare, "Raspberrypi models comparison — comparison tables," Disponível em: https://socialcompare.com/en/comparison/raspberrypi-models-comparison. Acesso em: 07 de julho 2022, 2022.

[21] S. Easterbrook, J. Singer, M.-A. Storey, and D. Damian, *Selecting Empirical Methods for Software Engineering Research*. London: Springer London, 2008, pp. 285–311.

[5] https://www.zabbix.com/