UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

GABRIEL DE OLIVEIRA MEIRA NOBREGA

PERFIS DE USO DE RECURSOS EM CLOUD

CAMPINA GRANDE - PB

2023

GABRIEL DE OLIVEIRA MEIRA NÓBREGA

PERFIS DE USO DE RECURSOS EM CLOUD

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel ou Bacharela em Ciência da Computação.

Orientador: Fábio Jorge Almeida Morais.

CAMPINA GRANDE - PB

2023

GABRIEL DE OLIVEIRA MEIRA NÓBREGA

PERFIS DE USO DE RECURSOS EM CLOUD

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel ou Bacharela em Ciência da Computação.

BANCA EXAMINADORA:

Professor Fábio Jorge Almeida Morais

Orientador – UASC/CEEI/UFCG

Professora Andrey Elisio Monteiro Brito

Examinador – UASC/CEEI/UFCG

Professor Tiago Lima Massoni

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 14 de fevereiro de 2023.

CAMPINA GRANDE - PB

RESUMO (ABSTRACT)[1]

With the advances on cloud it was never as fast to provision computing resources, that is due to the vast amount of resources these providers have. The cost of cloud applications can be very high, making the smart usage of these resources a crucial point for the industry today. To do that would allow for them to be even more accessible. The proposal for this article is to do an exploratory data analysis of an open source dataset provided by Azure, investigate different VMs profiles and get useful insights out of them.

1

Perfis de uso de recursos em cloud

Gabriel de Oliveira Meira Nóbrega
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
gabriel.nobrega@ccc.ufcg.edu.br

Fábio Jorge Almeida Morais
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
fabio@computacao.ufcg.edu.br

ABSTRACT

With the advances on cloud it was never as fast to provision computing resources, that is due to the vast amount of resources these providers have. The cost of cloud applications can be very high, making the smart usage of these resources a crucial point for the industry today. To do that would allow for them to be even more accessible. The proposal for this article is to do an exploratory data analysis of an open source dataset provided by Azure, investigate different VMs profiles and get useful insights out of them.

RESUMO

Com o advento da cloud nunca foi tão rápido o provisionamento de recursos computacionais, isso se dá pela vasta quantidade desses recursos que esses provedores possuem. O custo operacional de aplicações em cloud pode se tornar bastante elevado, tornando o uso inteligente desses recursos computacionais um ponto vital da indústria atualmente. Fazer isso pode permitir um barateamento dos recursos em geral e permitir negócios passem a ser viáveis, a proposta desse artigo é fazer uma análise exploratória de dados open source providos pela Azure e procurar investigar os diferentes perfis de uso nessa cloud e retirar insights do mesmo.

Keywords

Cloud,virtual machine, profiling, data analysis.

1. INTRODUÇÃO

De acordo com o relatório da Synergy Research Group (2022) o mercado de computação em nuvem vem crescendo a ritmos acelerados por anos, mantendo um crescimento de receita de 34% no ano de 2021. Isso se dá por conta da oferta de recursos computacionais sob demanda, que é altamente elástica e possui alta disponibilidade. Desta forma, torna-se extremamente útil para o ambiente de startups de rápido crescimento, pois o custo inicial para manter um datacenter *on-prem* acaba sendo evitado em prol de um custo operacional que pode rapidamente escalar de acordo com a demanda do produto.

Isso vem fazendo com que a computação seja novamente centralizada em grandes datacenters no lugar do on-prem que vem sendo deixado de lado em prol de uma computação mais centralizada [3]. Por isso é muito importante a utilização eficiente desses recursos, com o objetivo de minimizar custos de uso dos serviços de computação na nuvem. No entanto, a eficiência no uso desses recursos segue sendo um ponto importante pois pode criar um grande custo operacional se os recursos não forem utilizados de maneira inteligente. O intuito desse artigo é buscar padrões nesse uso e possíveis insights sobre como usar esses recursos de maneira mais eficiente..

Para isso, esse artigo faz uso de dados públicos disponibilizados pela Azure para investigar diferentes perfis de uso de VMs para avaliar as necessidades de cada um dos tipos de uso por meio de análises exploratórias de dados. Essa análise pode ajudar a melhor identificar o perfil dos usuários e entender como eles estão fazendo uso dos recursos provisionados para essas máquinas virtuais. Também pode ser útil para a Azure que de posse dessas informações pode melhor otimizar seus serviços a fim de se adequar melhor às necessidades dos seus usuários. Além disso, análises de dados como essa feita de maneira sucessiva também pode ser interessante a fim de identificar tendências de uso futuro desses usuários, dando uma vantagem estratégica para provedores que implementam essa estratégia.

2. TRABALHOS RELACIONADOS

No campo de perfilamento de VMs pode-se mencionar Naghmeh Dezhabad [7] que procura fazer uma análise de dados exploratória de dados publicamente disponibilizados pela Alibaba em instâncias spot para a utilização de batch jobs. O artigo conclui com uma divisão dessas demandas em três categorias, conseguindo dar um modelo de previsão de momentos de aumento de demanda para cada uma dessas categorias.

Já Z. Li [2] há um maior foco em identificar, selecionar e avaliar métricas para que seja objetivamente possível avaliar o desempenho das VMs dividindo as métricas em 3 categorias: Performance, Economia e Segurança. Porém conclui que a literatura atual foca bastante no quesito de performance e economia para produtos comerciais de cloud deixando de lado a Segurança.

Uma análise de dados sobre VMs em outro provedor, como a Azure, pode enriquecer a discussão no campo de perfilamento de VMs e ajudar na compreensão das tendências e padrões de uso em ambientes de nuvem. Sendo assim, essa análise pode ajudar a discussão acerca de estratégias de desenvolvimento de soluções de nuvem mais eficientes.

3. METODOLOGIA

Este artigo mostra os resultados de uma análise de dados exploratórias da base de dados públicos disponibilizados pelo provedor de computação na nuvem Azure, seguindo as seguintes etapas:

- Coleta e pré-processamento dos dados: Neste primeiro momento foi coletado e pré processado os dados a fim de eliminar dados inválidos e normalizar campos que não estejam adequadamente representados.

- Análise exploratória dos dados: Isso pode incluir tarefas como calcular estatísticas básicas, nesse caso se tratou de uma visualização da topografia dos dados de forma a identificar VMs comuns e padrões de uso mais claros.

- Análise dos dados para responder às perguntas de pesquisa: A partir dessas perguntas foi realizada uma análise de dados mais objetiva a fim de identificar possíveis respostas para essas perguntas.

A partir dos dados disponibilizados pela Azure no repositório com os dados públicos em [1], vamos realizar um tratamento e análises exploratórias. Se tratam de dados coletados de 2,695,548 VMs num período de 30 dias contendo os seguintes dados:

- vm id(id da VM),
- timestamp vm created(timestamp da criação da VM),
- timestamp vm deleted(timestamp da finalização da VM),
- max cpu(uso máximo da CPU em percentagem)
- avg cpu(uso médio da CPU em percentagem)
- p95 max cpu(95^{o} percentil de uso da CPU em percentagem)
- vm category(categoria da VM)
- vm virtual core count bucket(número de cores da VM)
- vm memory (gb) bucket(quantidade de RAM em gigas)

Propomos usar esses dados para fazer análises exploratórias e perfilamento das VMs afim de procurar identificar perfis dessas VMs, após isso irá ser avaliado como cada VM utilizou os recursos provisionados para a mesma em termos de questões de uso de cpu, memória e duração da vm.

Para isso se fez necessário a criação de novos campos, como duração e tratamento de dados, no caso de VMs com mais de 30 cores foram todas transformadas em apenas 30 cores. A mesma lógica foi utilizada no tratamento dos dados referente à quantidade de memória ram disponíveis para as VMs.

4. RESULTADOS

Ao investigar a distribuição das categorias das VMs foi observado que 91% das VMs são do tipo Unknown como é evidenciado pela Figura 1, que mostra a quantidade de VMs de cada tipo presente nos dados. No geral, VMs do grupo Unknown tendem a ter menor duração (tempo de execução_, isso pode indicar que essas VMs desse grupo são batch jobs gerais que atendem muitos serviços do Azure. Isso pode ser observado na Figura 2, que mostra a relação entre o tipo da VM, o uso máximo de cpu e a duração da VM.

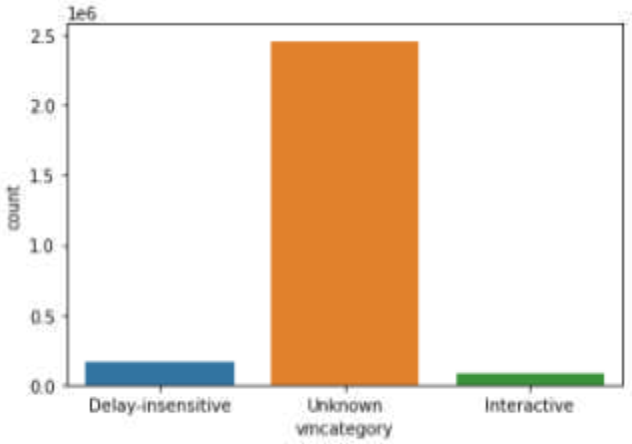

Fig. 1: Distribuição categorias das vms

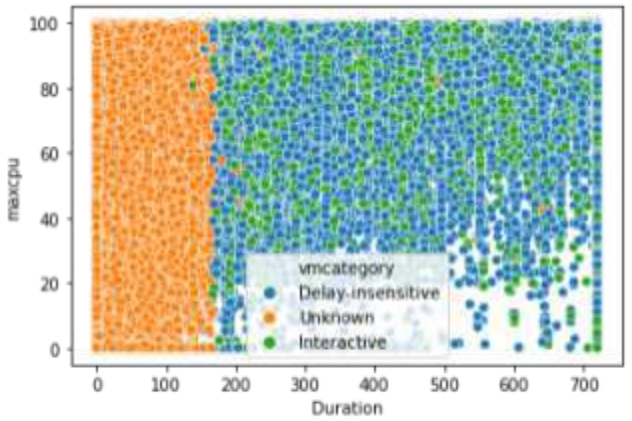

Fig. 2: Scatter plot das vms do uso maximo em relação ao tempo

Com o intuito de verificar que tipo de VMs costuma ser mais utilizado foi avaliado a distribuição dessas VMs em termos da memória e do números de cores provisionados. Através disso foi observado que há uma grande concentração de VMs com poucas cores, diminuindo em quantidade conforme as VMs aumentam de

tamanho, como pode ser visto na Figura 3.

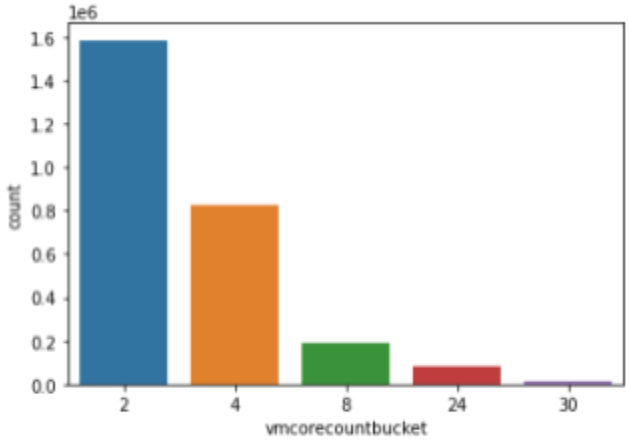

Fig 3: Número de vms para cada número de núcleos

A distribuição de memória está apresentada na Figura 4, essa maior concentração em VMs com 8 GB e 32 GB pode implicar uma maior demanda para VMs cujo workload seriam mais *memory heavy*.

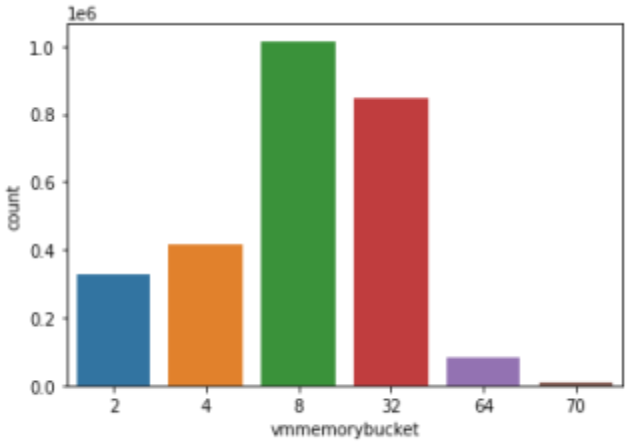

Fig 4: Barplots com distribuição das memórias

As VMs possuem normalmente uma quantidade proporcional de cores e memória, são poucas as com grande quantidade de memória e pouco processamento e nenhuma com muito processamento e pouca memória, vide Figura 5.

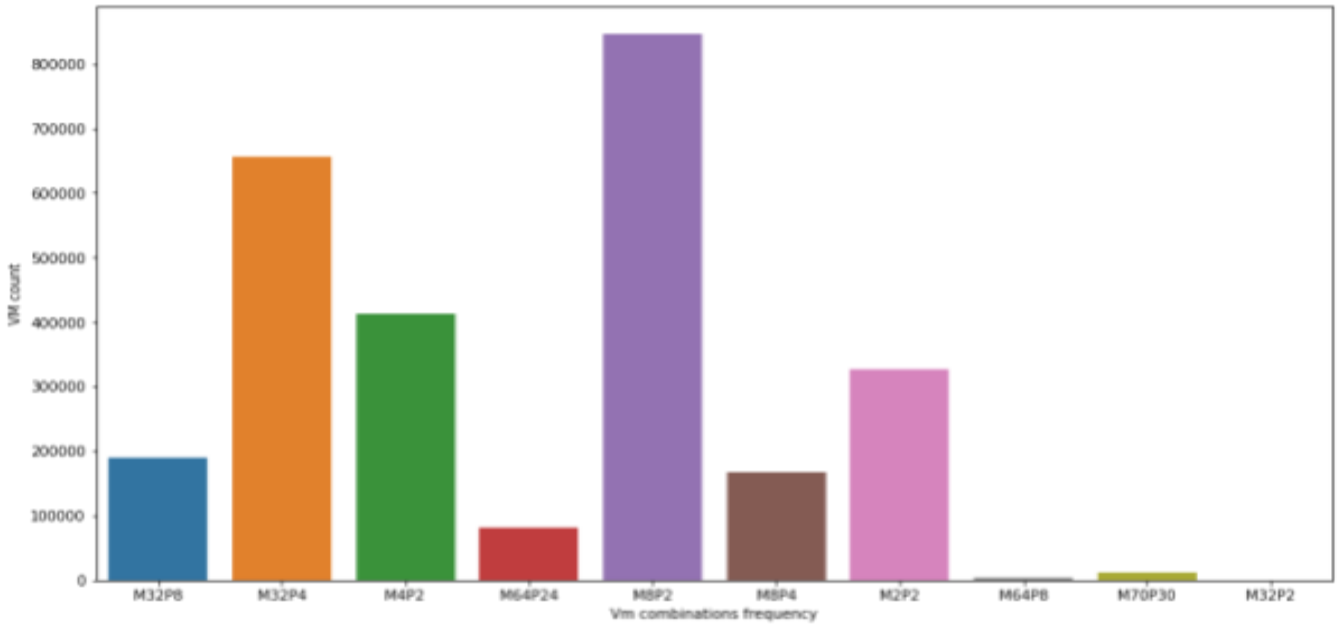

Fig. 5: Configurações da vm em ram e número de núcleos

Porém, ao realizar o cruzamento desses dados do número de cores e o aproveitamento deste processamento, foi identificado uma certa correlação entre o uso máximo de CPU e o número de *cores* disponíveis conforme visto na Figura 6. Ou seja, quanto maior o número de cores alocado, maior é a utilização máxima média. Isso pode indicar um melhor aproveitamento dos recursos provisionados para essa categoria de VM [2].

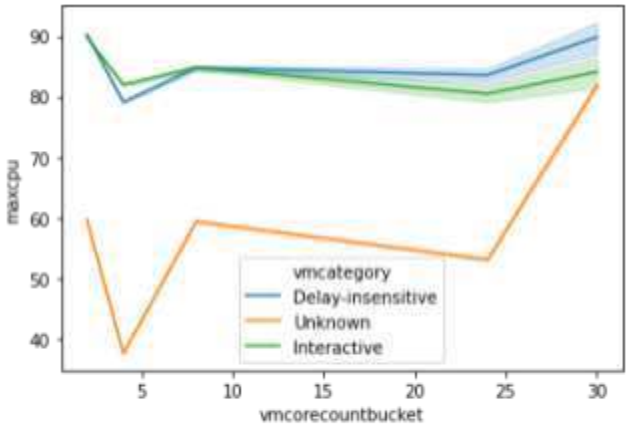

Fig 6: Lineplot para o uso de CPU máximo em relação ao número de cores.

Essa tendência também é possível de ser evidenciada quando se observa os boxplots do uso médio e máximo de CPU para VMs com diferentes quantidade de cores alocadas. Para VMs com a alocação de 30 cores (Figura 7) o uso máximo de CPU se aproxima dos 100%, enquanto que para VMs com menos cores a utilização máxima é menor e mais variada (Figura 8).

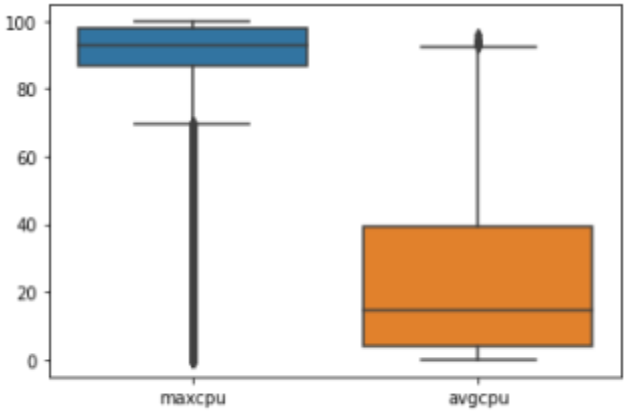

Fig 7: Boxplot do uso de CPU para 30 cores.

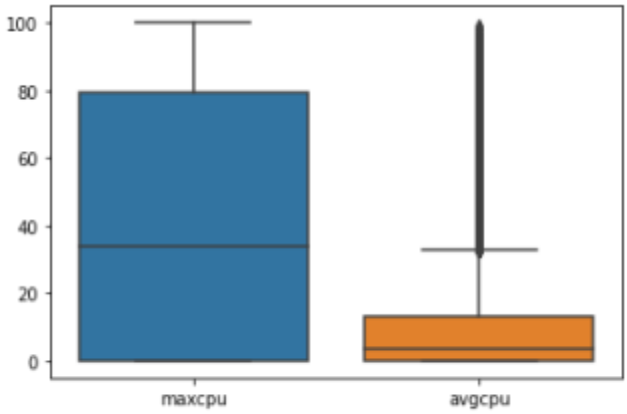

Fig 8: Boxplot do uso de CPU para 4 cores

Isso pode ser decorrente de vários fatores, dentre eles o fato dessas VMs serem mais caras e de uso mais específico, o que pode levar a um provisionamento mais criterioso e por isso um uso mais adequado aos requisitos do projeto.

Foi observado que independente do números de cores, as VMs com categoria determinada (*Delay-insensitive* e *Interactive)* possuíam uma tempo maior de uso. A Figura 9 mostra a duração de uso das VMs para diferentes categorias.

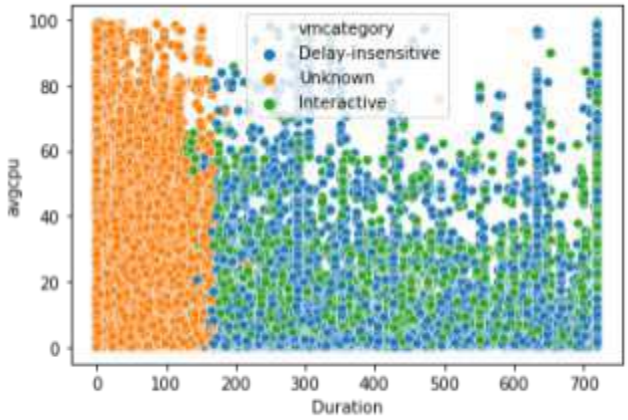

Fig 9: VM's e suas categoria no mesmo gráfico

Procurando agrupar essas VMs com base nas suas características foi utilizado um algoritmo *KMeans* para tentar identificar esses clusters, em termos do uso médio de CPU e da duração das VMs. A Figura 10 mostra o resultado desse agrupamento para 5 grupos. Com base nesses resultados, foi verificado que VMs com menor duração (menos de 100 minutos, todas Unknown (Figura 9), provavelmente se tratando de *batch jobs*) estão divididas em dois grupos, um com maior utilização média de CPU (em vermelho) e outro com menor utilização(em azul). Os demais grupos possuem VMs de todas as categorias, com uma maior predominância das VMs com categoria definida, que são agrupados com base na duração da VM.

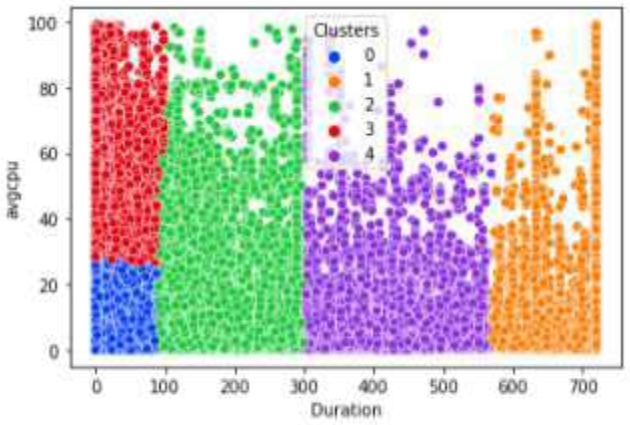

Fig 10: Clusterização das VM utilizando 5 nós kmeans

4. CONCLUSÃO

Nesse artigo foi possível realizar uma análise das características de VMs da Azure, com base em um conjunto de dados públicos disponibilizado publicamente na internet. Para isso foi necessário o uso de diversas técnicas de análise de dados com o intuito de caracterizar essas máquinas. Nesse âmbito de máquinas do Azure foi identificado que a maior parcela de VMs são do tipo Unknown e que VMs com mais cores possuem em geral uma utilização maior de recursos de CPU.

A área de métricas de avaliação de desempenho segue sendo um tema essencial para avaliar sistemas de computação, devido a falta de consenso da área na definição dessas. Isso fica ainda mais evidente em sistemas diversos que podem estar cumprindo diferentes funções que pode ser o caso em IaaS como as VMs estudadas. Por isso há espaço para pesquisas que procuram delimitar tais métricas.

Este trabalho foi feito no contexto de um Trabalho de Conclusão de Curso de Ciência da Computação com o intuito de avaliar e perfilar o uso de recursos de VM´s da Azure.

5. REFERÊNCIAS

[1] Mohammad Shahrad, Rodrigo Fonseca, Inigo Goiri, Gohar Chaudhry, Paul Batum, Jason Cooke, Eduardo Laureano, Colby Tresness, Mark Russinovich, Ricardo Bianchini. "Serverless in the Wild: Characterizing and Optimizing the Serverless Workload at a Large Cloud Provider", in Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC 20). USENIX Association, Boston, MA, July 2020.

[2] Z. Li, L. O'Brien, H. Zhang and R. Cai, "On a Catalogue of Metrics for Evaluating Commercial Cloud Services," Proc. of the 13th International Conf. on Grid Computing, 2012, pp.164-173

[3] Ajeh, Daniel, Ellman, Jeremy and Keogh, Shelagh (2014) A cost modelling system for cloud computing. In: The 14th International Conference on Computational Science and Applications (ICCSA 2014), 30 June -3 July 2014, University of Minho, Guimaraes, Portugal.

[4] Blesson Varghese, Ozgur Akgun, Ian Miguel, Long Thai and Adam Barker(2014) Cloud Benchmarking for Performance. In: CloudCom - IEEE International Conference on Cloud Computing Technology and Science

[5] Daniel Ajeh, Jeremy Ellman and Shelagh Keogh(2014) A Cost Modelling System for Cloud Computing. In: The 14th International Conference on Computational Science and Applications (ICCSA 2014)

[6] Joel Scheuner e Philipp Leitner(2018) Estimating Cloud Application Performance Based on Micro-Benchmark Profiling. In 2018 IEEE 11th International Conference on Cloud Computing (CLOUD)

[7] Naghmeh Dezhabad, Sudhakar Ganti and Gholamali Shoja(2019) Cloud Workload Characterization and Profiling for Resource Allocation. In: 2019 IEEE 8th International Conference on Cloud Networking (CloudNet)

[8] Dimiter R. Avresky, Pierangelo Di Sanzo, Alessandro Pellegrini, Bruno Ciciani and Luca Forte(2014) Proactive Cloud Management for Highly Heterogeneous Multi-Cloud Infrastructures. In: Institute of Electrical and Electronics Engineers (IEEE)

[9] RENO, NV, Huge Cloud Market Still Growing at 34% Per Year; Amazon, Microsoft & Google Now Account for 65% of the Total em:https://www.srgresearch.com/articles/huge-cloud-market-is-still-growing-at-34-per-year-amazon-microsoft-and-google-now-account-for-65-of-all-cloud-revenues. Acesso em 31 jan 2023