Critérios para adoção de computação em nuvem para execução de técnicas de aprendizado de máquina

Adriana Melges Quintanilha Weingart[1]*; Prof. Dr. Thiago Bianchi[2]

[1] Kyndryl Brasil Serviços Ltda. IT Architect Senior, Distinguished Technical Specialist. São Paulo, Brasil.
[2] Itaú Unibanco. Gerente de Ciência de Dados. São Paulo, Brasil.

Critérios para adoção de computação em nuvem para execução de técnicas de aprendizado de máquina

Resumo

Dados estão em todo lugar e fazem parte da vida de todas as pessoas. A análise de dados não está mais restrita às empresas, mas profissionais independentes e pessoas com interesses em determinados assuntos também buscam extrair informações deles. Com esse grupo de pessoas em mente, este estudo buscou identificar e comparar critérios para a escolha do recurso para análise dos dados entre equipamentos pessoais ou recursos de "cloud" de forma a permitir uma decisão informada baseada em escolhas e necessidades individuais. Usando algoritmos de análise não-supervisionada na análise de dados abertos do ensino médio do Brasil para teste, este estudo gerou alguns cenários que foram executados em diferentes ambientes. A comparação dos resultados destes cenários de teste forneceu os elementos necessários para a análise proposta e a identificação de melhores usos de cada cenário.

Palavras-chave: cloud computing; cloud; ENEM; comparação; análise de agrupamentos; análise fatorial por componentes principais; PCA; análise de correspondência múltipla; ACM

Criteria for adopting cloud computing for executing machine learning techniques

Abstract

Data is everywhere and it is part of everyone's life. Data analysis is no longer restricted to large companies, but independent professionals and people with interest in different subjects also want and try to extract information from it. With this group in mind, this study sought to identify and compare criterias for choosing the equipment for data analysis between personal computer or cloud resources to allow an informed decision based on individual choices and needs. Using unsupervised algorithms in the analysis of open data from the Brazilian high school for testing, this study generated some scenarios that were executed in different environments. These tests scenarios' results comparison provided the necessary elements for the proposed analysis and the identification of best uses depending on the situation.

Keywords: cloud computing; cloud; ENEM; comparison; cluster analysis; principal component factor analysis; correspondence analysis

Introdução

O registro de dados é uma característica do ser humano desde a antiguidade, e a cada era que passa mais dados são criados e armazenados. E este processo está acelerando cada vez mais. De acordo com Vopson (2021), em 2018 foi gerado e armazenado em todo o mundo um total de 33ZB (zettabytes), com este valor crescendo para 59ZB em 2020. A projeção para 2025 é chegar a 175ZB de dados gerados, capturados e armazenados.

Mas um dado simplesmente armazenado não traz valor para quem o armazena ou para a comunidade. Os dados precisam ser compartilhados, e o mais importante, analisados para que possam produzir informações importantes e relevantes para uma empresa, uma comunidade, para a humanidade. Atualmente, além das empresas buscarem informações relevantes para seus negócios e estratégia, muitas pessoas - consultores, pesquisadores

independentes, professores e alunos ou simplesmente curiosos - encontram nesta imensidão de dados uma oportunidade para explorar e extrair informações, seja para resolver um problema, seja para suportar uma pesquisa, ou apenas para satisfazer uma curiosidade.

Fávero e Belfiori (2017) ressaltam características de geração e disponibilidade de dados que marcam os dias atuais: volume, velocidade, variedade, variabilidade e complexidade. Estas características definem “Big Data”: dados são gerados em volume e velocidade extraordinários, com uma variedade, variabilidade e complexidade crescentes. E isso torna a análise manual dos dados inviável, se não impossível.

Neste contexto surge outro importante componente: a Computação em Nuvem, do inglês “Cloud Computing”. O “National Institute of Standards and Technology [NIST]” define computação em nuvem como a habilitação de recursos de computação, armazenamento e processamento de dados, sob demanda, rapidamente, com mínimo esforço ou interação humana e acessível de qualquer lugar e a qualquer momento. Essas características da computação em nuvem facilitam a geração e crescimento exponencial dos dados, seu armazenamento e sua análise.

Assim, associando essa necessidade do ser humano de armazenar, analisar e entender os dados para extrair informações úteis, à tradicional estatística e ao novo poder computacional proporcionado pela computação em nuvem, novos métodos e algoritmos são criados, surgindo assim a Ciência de Dados, que nada mais é que a consolidação de elementos variados, baseada em técnicas e teorias vindas de outros campos básicos do conhecimento humano, como matemática, estatística, engenharia e ciências (Porto e Ziviani, 2014).

Toda essa evolução preparou o cenário atual da humanidade: grandes volumes de dados gerados constantemente, podendo ser armazenados e processados de qualquer lugar e a qualquer momento, com grande poder de processamento e custo consideravelmente menor, e a análise destes dados para extração de informações úteis e relevantes.

Muitos provedores de serviço de “cloud computing” (“Cloud Service Provider [CSP]”) estão disponíveis no mercado, cada um com propostas e planos diferentes e com foco em nichos distintos de mercado. Muitos algoritmos foram criados em várias linguagens de programação (sem mencionar os programas e recursos dedicados de nuvem) para análise dos dados.

E essas possibilidades abrem espaço para novos questionamentos: como escolher o melhor serviço de nuvem? O que considerar em uma escolha? É caro? É realmente mais performático? É opção para pequenas empresas, profissionais independentes e curiosos que querem trabalhar e entender dados?

Este estudo busca identificar, através de pesquisa em estudos acadêmicos e de mercado, alguns principais fatores para a escolha de um CSP e, a partir desta identificação, realizar testes, levantamento e análise de dados de alguns provedores disponíveis no mercado de computação em nuvem para mapear critérios para escolha. Espera-se que os resultados deste estudo possam auxiliar profissionais independentes, como consultores, autônomos, professores e alunos, curiosos e apaixonados por dados, estendendo-se para pequenas e médias empresas, na escolha informada de um ambiente adequado para suas necessidades e seus trabalhos de pesquisa.

Material e Métodos

Para atingir o objetivo do estudo, de identificar os principais fatores para escolha de provedor de serviços de “cloud computing” e, a partir desta escolha, mapear critérios a serem considerados para a mesma, este trabalho teve um caráter qualitativo e quantitativo no que diz respeito à sua abordagem. Qualitativo ao preocupar-se com aspectos da escolha que não podem ser quantificáveis, mas baseando-se, como mencionado por Gerhardt e Silveira (2009), em objetivação do fenômeno e compreensão, explicação e observância das orientações teóricas e dados empíricos. Por outro lado, alguns critérios podem ser quantificáveis, o que permitiu que a pesquisa tivesse também um caráter prático e exploratório ao utilizar uma abordagem quantitativa, com pesquisa experimental para coleta de dados, análise dos dados brutos extraídos com padronização e neutralidade durante a execução dos testes e recorrendo à matemática para estabelecimento de relações e análises (Gerhardt e Silveira, 2009).

Já quanto à natureza, ela teve um caráter de pesquisa prática uma vez que as informações produzidas permitem aplicação prática e focada a uma comunidade específica.

Por ter como objetivo a exploração de dados de execução de análise (Gerhardt e Silveira, 2009), ela teve caráter de pesquisa exploratória.

E, finalmente, quanto ao procedimento, esta pesquisa foi experimenta, seguindo os seguintes critérios:

Definição dos CSP para avaliação

Atualmente, há disponíveis no mercado vários provedores de serviço de cloud (“Cloud Service Provider [CSP]”), com recursos semelhantes e custos concorrentes. Como o teste em todos estes ambientes não era viável, este estudo considerou dois critérios para escolha, ambos baseados em pesquisas de 2021 do Gartner, uma empresa de consultoria e pesquisa tecnológica fundada em 1979 e baseada em Stamford, Connecticut, EUA. As pesquisas

consideradas foram: “Market Share”[1] e Posicionamento no Quadrante Mágico[2] de Visão por habilidade de executar.

A análise de dados pode utilizar tanto “Infrastructure as a Service [IaaS]”, ou seja, recursos de infraestrutura, como servidores e máquinas virtuais, “storage” e rede, como “Platform as a Service [PaaS]”, que inclui serviços de plataforma, além dos listados em IaaS, permitindo menor preocupação com infraestrutura e maior foco no desenvolvimento das aplicações. Há ainda a possibilidade de uso de serviços “Software as a Service [SaaS]”, no qual a pessoa tem a mínima interação com a infraestrutura do CSP, ao mesmo tempo que pode usar a aplicação disponibilizada normalmente.

Como a classificação do “Market Share” é somente para IaaS e o Quadrante Mágico considera os dois (IaaS e PaaS) em sua análise, o segundo reflete melhor os principais CSP no mercado e foi considerado para este estudo, sendo escolhidos Amazon Web Services [AWS], Microsoft Azure [Azure] e Google Cloud Platform [GCP].

Adicionalmente, outro serviço foi incluído no estudo, para explorar outra opção de provisionamento, codificação e testes iniciais, na modalidade gratuita: RStudio Cloud.

É importante ressaltar que este estudo não tem objetivo de escolher um provedor de serviço de nuvem, mas identificar critérios para escolha.

Definição dos critérios para análise e comparação

Para a escolha dos critérios de análise e comparação, este estudo considerou as conclusões dos estudos realizados por Lang et al. (2018) e Alabool et al. (2018).

Lang et al. (2018), em seu estudo usando metodologia Delphi e diversos profissionais, identificou 31 atributos de qualidade de serviço para seleção de provedores de serviços de “cloud”. Alabool et al. (2018), por sua vez, mapeou os temas mais abordados em mais de 77 artigos publicados, identificando como critérios mais usados: custo, segurança e performance.

Como segurança em nuvem deve adotar um conceito de responsabilidade compartilhada (Lane et al., 2017), dependendo de definições e usos seguro das pessoas envolvidas e não somente do CSP, este critério não foi considerado para análise neste estudo.

[1] Magic Quadrant for Cloud Infrastructure and Platform Services
https://www.gartner.com/doc/reprints?id=1-271OE4VR&ct=210802&st=sb

[2] Magic Quadrant for Cloud Infrastructure and Platform Services
https://aws.amazon.com/resources/analyst-reports/gartner-mq-cips-2021/

Definição dos ambientes de testes

A linguagem escolhida para este estudo foi R. Como mencionado por Kumari e Verma (2019), o R é uma linguagem bem estruturada para dados, avaliação de fatos e bons algoritmos para informação e mineração de dados, permitindo uma grande variedade de técnicas através de seus pacotes/algoritmos. Há também que se considerar que o R é uma linguagem "open-source" e independe de plataforma. Mesmo uma de suas desvantagens, o armazenamento dos objetos na memória física que pode impactar na performance e tempo de reposta, contribui para este estudo ao permitir avaliar cenários com melhor desempenho.

A codificação e testes iniciais foram realizados em computadores pessoais ("Personal Computer [PC]"), fora da nuvem, como descrito na Tabela 1.

Tabela 1. Detalhamento do ambiente de codificação e testes iniciais

#	Equipamento / instância	SO(1) e versão	Proc(2), Mem(3), Disco	R	RStudio	Custo
A01	MacBook Pro 2017(4)	macOS Monterey, versão 12.5	Dual Core, 8GB, 256GB	4.2.1	2022.07.1 build 554	R$15.300,00(5)
A02	DELL Vostro 361	Windows 10 Pro	i7-10700 (8-core), 8GB, 512GB	4.2.1	2022.07.2 build 576	R$5.800,00(6)

Fonte: Dados originais da pesquisa
Nota: (1) Sistema Operacional [SO]; (2) Processador [Proc]; (3) Memória [Mem]; (4) Especificações técnicas do MacBook Pro - https://support.apple.com/kb/SP754?locale=pt_PT; (5) Valor de um equipamento equivalente novo, no site do fabricante (https://www.apple.com/br/shop/buy-mac/macbook-pro/13-polegadas); (6) Valor do equipamento em novembro 2021

Apesar dos CSP terem diversas opções para Ciência de Dados e "Machine Learning", este estudo optou por usar o RStudio. Essa abordagem permitiu a comparação dos resultados com a execução dos mesmos testes em equipamentos pessoais.

A versão do R utilizada dependeu das disponíveis para instalação ou provisionamento nos ambientes de nuvem em estudo. Se nenhuma imagem estava disponível pelo provedor, foi utilizada a última versão do software R, baixada do "Comprehensive R Archive Network [CRAN]", pelo link https://www.r-project.org. A mesma variação ocorreu para Sistema Operacional [SO] e capacidade (memória e CPU) do ambiente.

Inicialmente, foi dado preferência às opções gratuitas dos serviços em suas configurações disponíveis, que foram reconfigurados para maior capacidade para os testes seguintes. A escolha da região do provedor de nuvem foi feita baseada no menor custo (e não distância, uma vez que latência na transmissão dos dados não é um critério de avaliação neste estudo).

A Tabela 2 contém as configurações usadas em cada ambiente testado. É importante ressaltar que os custos documentados na tabela são da máquina virtual ("Virtual Machine

[VM]") escolhida, e outros custos de provisionamento na nuvem podem incorrer, como discos e egresso dos dados, dentre outros.

Tabela 2. Detalhamento dos ambientes de testes

#	Provedor do serviço[6]	Modelo de serviço	SO[1] e versão	Proc[2], Mem[3], Disco	R	RStudio	País e Região, Custo / h
A03	CSP02	IaaS	Ubuntu 18.04LTS	1 vCPU, 1GiB, 30GiB	4.0.2	Server 1.3.1073	US East, n/a
A04	RStudio Cloud	SaaS[4]	Ubuntu 20.04	1 CPU, 1GB, n/a	4.2.1	2022.02.2 build 485	n/a
A05	CSP01	IaaS	Ubuntu 18.04LTS	2 vCPU, 8GB, 10GB	4.2.1	2022.07.1 build 554	US Central, US$0,07
A06	CSP01	IaaS	Ubuntu 18.04LTS	2 vCPU, 8GB, 70GB (SSD)[5]	4.2.1	2022.07.1 build 554	US Central, US$0,08
A07	CSP01	IaaS	Ubuntu 18.04LTS	8 vCPU, 32GB, 70GB (SSD)	4.2.1	2022.07.1 build 554	US Central, US$0,28
A08	CSP02	IaaS	Ubuntu 18.04LTS	2 vCPU, 8GB, 30GB	4.0.2	Server 1.3.1073	US East, US$0,0928
A09	CSP03	IaaS	Ubuntu 18.04 Gen2	2 vCPU, 8GB, n/a	3.4.4	Server 1.3.1093	US West, US$0,086
A10	CSP03	IaaS	Ubuntu 18.04 Gen2	4 vCPU, 16GB, n/a	3.4.4	Server 1.3.1093	US West, US$0,17
A11	CSP03	IaaS	Ubuntu 18.04 Gen2	4 vCPU, 32GB, n/a	3.4.4	Server 1.3.1093	US West, US$0,23
A12	CSP02	IaaS	Ubuntu 18.04LTS	8 vCPU, 32GB, 30GB	4.0.2	Server 1.3.1073	US East, US$0,23

Fonte: Dados originais da pesquisa
Nota: [1] Sistema Operacional [SO]; [2] Processador [Proc]; [3] Memória [Mem]; [4] "Software as a Service [SaaS]"; [5] "Solid-state drive [SSD]"; [6] Os nomes dos provedores de nuvem foram omitidos para evitar vieses

As instâncias menores nos ambientes de nuvem foram provisionadas seguindo os procedimentos criados e disponibilizados no GitHub, acessível pelo link https://bit.ly/AW_TCC_Procedimentos:

- Google Cloud - Criando uma instalação do RStudio.pdf
- AWS - Criando uma instalação do RStudio.pdf
- RStudio Cloud - Criando uma instância.pdf
- Azure - Criando uma instalação do RStudio.pdf

Uma vez feito o teste no ambiente com capacidade menor, ao invés de fazer novo provisionamento de um ambiente maior, foi feito um "upgrade" do ambiente, tendo em consideração a elasticidade permitida no ambiente de nuvem, ou seja, a habilidade de aumentar e reduzir capacidade rapidamente, sob demanda e automaticamente, como explicado por Herbst et al. (2013).

A última versão do software R foi baixado do CRAN para os PCs, escolhendo um "mirror" da Universidade de São Paulo [USP], Piracicaba e instalado com as opções padrões. De maneira similar, a última versão do RStudio[3] foi baixada do site https://rstudio.com e instalada com as opções padrões.

[3] Há uma nota no site do RStudio informando que a partir de outubro 2022, o RStudio passará a se chamar Posit - https://rstudio.cloud/learn/posit

Definição dos dados e algoritmos para testes

Para poder avaliar diversos aspectos, diferentes modelos foram definidos para teste. Cada modelo foi testado três vezes, em momentos diferentes, em todos os ambientes de teste definidos. A média das execuções foi considerada para análise.

Os modelos usados para este estudo estão listados na Tabela 3.

Tabela 3. Cenários de testes usados

#	Amostra	Modelo	Pacotes usados
T1	Amostra normal gerada	Modelo de regressão simples	stats, tictoc
T2	Microdados do ENEM 2020(1)	Análise Exploratória dos dados	tidyverse, stringi, gridExtra, kableExtra, psych, tictoc, knitr
T3	Microdados do ENEM 2020(1)	Análise de Agrupamentos e Análise de Correspondência Múltipla [ACM]	tidyverse, kableExtra, knitr, tictoc, factoextra, cabootcrs, FactoMineR, ggrepel, gridExtra
T3r	Microdados do ENEM 2020(1) do Estado de Roraima	Análise de Agrupamentos e Análise de Correspondência Múltipla [ACM]	tidyverse, kableExtra, knitr, tictoc, factoextra, cabootcrs, FactoMineR, ggrepel, gridExtra
T4	Microdados do ENEM 2020(1)	Análise Fatorial por Componentes Principais [PCA]	tidyverse, kableExtra, knitr, tictoc, reshape2, psych, PerformanceAnalytics, ggrepel, plotly, factoextra
T4r	Microdados do ENEM 2020(1) do Estado de Roraima	Análise Fatorial por Componentes Principais [PCA]	tidyverse, kableExtra, knitr, tictoc, reshape2, psych, PerformanceAnalytics, ggrepel, plotly, factoextra

Fonte: Dados originais da pesquisa
Nota: (1) Dados do Exame Nacional do Ensino Médio [ENEM] disponíveis no site do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP] em https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem.

A amostra normal do teste T1, com 5.000.000 elementos, foi criada com uma função do R que gera desvios aleatórios para a distribuição normal.

Para os testes T2, T3 e T4 foram utilizados os microdados do Exame Nacional do Ensino Médio [ENEM] de 2020, cujo arquivo no formato “Comma-separated values [CSV]” foi disponibilizado ao público pelo site do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP]. Esse arquivo de 2,02GB contém um “dataset” com 76 variáveis para 5.783.109 observações. Como a quantidade de observações era muito grande, a amostra foi reduzida para processamento e análise inicial considerando o Estado Brasileiro com menos

entradas, Roraima [RR], e as versões T3 reduzida (T3r) e T4 reduzida (T4r) foram usadas para codificação e testes iniciais. Ao final, a versão T3r foi incluída em todos os testes também.

O código do teste T1 foi feito em formato RScript, enquanto o código dos testes T2, T3/T3r e T4 foram feitos em formato R Markdown. Todos os códigos utilizados nos cenários de teste foram disponibilizados no GitHub, acessível pelo link https://bit.ly/AW_TCC_Scripts.

Os arquivos em formato "HyperText Markup Language [HTML]" gerados da execução dos códigos em R Markdown também foram disponibilizados no GitHub, acessível pelo link https://bit.ly/AW_TCC_ENEM-2020.

Cada um destes cenários foi subdivido em etapas, para explorar também as diferenças no processamento e algoritmos, como detalhado nas Tabela 4.

Tabela 4. Detalhes do cenário de testes T1 – Amostra normal gerada

(continua)

Teste	Etapa	Processamentos
T1	Geração dos Dados	Geração dos dados, através da função rnorm() para análise/teste
T1	Execução do Modelo	Execução do modelo - Regressão simples
T2	Inicialização	Carregamento dos pacotes do R e leitura do "dataset"
T2	Tratamento dos Dados	Limpeza dos dados (participantes que faltaram ou zeraram em alguma prova e entradas NA) e remoção de variáveis que não foram utilizadas no estudo
T2	Execução do Modelo	Distribuição dos participantes por Estado brasileiro, desempenho dos participantes nas provas de Matemática e Redação e comparação dos dados e desempenho por tipo de escola
T3/T3r	Inicialização	Carregamento dos pacotes do R e leitura do "dataset"
T3/T3r	Tratamento dos Dados	Limpeza dos dados (participantes que faltaram ou zeraram em alguma prova e entradas NA) e remoção de variáveis que não foram utilizadas no estudo – no caso do teste reduzido, o "dataset" também foi reduzido nesta etapa a somente os dados de Roraima
T3/T3r	Análise de Agrupamentos	Definição do "cluster" (pelos métodos "elbow", "silhouette" e "gap statistic method"), cálculo dos agrupamentos e análise
T3/T3r	ACM	QUI2 Test, Matrizes binária e de Burt, ACM e Mapa Perceptual, e Análise dos agrupamentos com os resultados da ACM
T4	Inicialização	Carregamento dos pacotes do R e leitura do "dataset"

Tabela 4. Detalhes do cenário de testes T1 – Amostra normal gerada

(conclusão)

Teste	Etapa	Processamentos
T4	Tratamento dos Dados	Limpeza dos dados (participantes que faltaram ou zeraram em alguma prova e entradas NA) e remoção de variáveis que não foram utilizadas no estudo – no caso do teste reduzido, o "dataset" também foi reduzido nesta etapa a somente os dados de Roraima.
T4	PCA	Avaliação do uso, matriz de correlações, padronização dos dados, PCA, definição dos fatores e cargas fatoriais, e construção de um "ranking" de classificação dos participantes

Fonte: Dados originais da pesquisa

Definição das ferramentas de visualização dos dados para análise

O MS Excel foi utilizado tanto para a tabulação dos resultados e para a criação de gráficos para análise visual.

Os gráficos gerados foram revisados para garantir acessibilidade (no caso de pessoas com daltonismo), no site Colblindor[4], através da ferramenta online gratuita Coblis – "Color Blindness Simulator"[5].

Resultados e Discussão

Esta sessão traz os resultados da execução e análise de cada cenário de teste proposto nos ambientes definidos.

Cenário de teste T1 – Modelo de Regressão Simples em amostra normal

A Tabela 5 apresenta os resultados obtidos nas diversas execuções do teste T1 (regressão simples em amostra normal). Pode-se observar que a fase Execução do Modelo falhou no ambiente A04 (RStudio, SaaS), com uma mensagem genérica de erro inesperado. Por ser um ambiente SaaS, a análise da causa do erro é limitada e depende do CSP e isso baseou a decisão de não incluir este ambiente nos demais testes.

[4] Colblindor: https://www.color-blindness.com/
[5] Coblis: https://www.color-blindness.com/coblis-color-blindness-simulator/

Tabela 5. Resultados das execuções do teste T1 nos ambientes definidos para estudo

Ambiente	Execução	Geração dos dados(1)	Execução do Modelo(1)	Tempo Total(1)
A01	#1	4,142	120,276	124,418
A01	#2	5,947	135,955	141,902
A01	#3	4,100	118,035	122,135
A02	#1	2,940	40,060	43,000
A02	#2	2,780	41,090	43,870
A02	#3	2,640	42,080	44,720
A03	#1	5,350	186,690	192,040
A03	#2	7,150	198,920	206,070
A03	#3	5,800	213,990	219,790
A04	#1	3,475	falha / erro	-
A04	#2	5,005	falha / erro	-
A04	#3	4,750	falha / erro	-
A05	#1	4,776	63,602	68,378
A05	#2	5,601	64,222	69,823
A05	#3	4,503	62,342	66,845
A06	#1	5,321	37,704	43,025
A06	#2	4,719	63,601	68,320
A06	#3	4,760	67,465	72,225
A07	#1	4,625	62,468	67,093
A07	#2	5,497	63,006	68,503
A07	#3	3,876	61,766	65,642
A08	#1	4,705	66,928	71,633
A08	#2	4,255	50,317	54,572
A08	#3	4,870	50,565	55,435
A09	#1	2,168	34,615	36,783
A09	#2	2,236	35,079	37,315
A09	#3	2,049	34,613	36,662
A10	#1	2,121	33,628	35,749
A10	#2	1,915	33,335	35,250
A10	#3	1,991	33,867	35,858
A11	#1	2,184	34,717	36,901
A11	#2	2,024	34,336	36,360
A11	#3	2,179	34,705	36,884
A12	#1	4,449	51,472	55,921
A12	#2	4,030	53,131	57,161
A12	#3	3,862	43,815	47,677

Fonte: Resultados originais da pesquisa
Nota: (1) Valores em segundos

Para facilitar o entendimento e comparação dos resultados, as médias dos tempos de execução da etapa Geração de Dados podem ser visualizadas graficamente na Figura 1. Da comparação com os PCs (A01 e A02), pode-se observar que:

- os ambientes em nuvem A04, A05, A06, A07, A08 e A12, com tempos médios de execução similares entre si e ao PC de A01, apesar de terem versões mais recentes do software R, tem configurações diferentes de infraestrutura. Isso indica que a quantidade de memória ou de CPU de cada ambiente não afetou tempo de processamento.

- os ambientes em nuvem A02, A09, A10 e A11 tiveram os melhores tempos de processamento, sendo que os três últimos apesar de configurações de capacidade diferentes e versão mais antiga do software R, estão no mesmo provedor de nuvem e tiveram os melhores tempos.

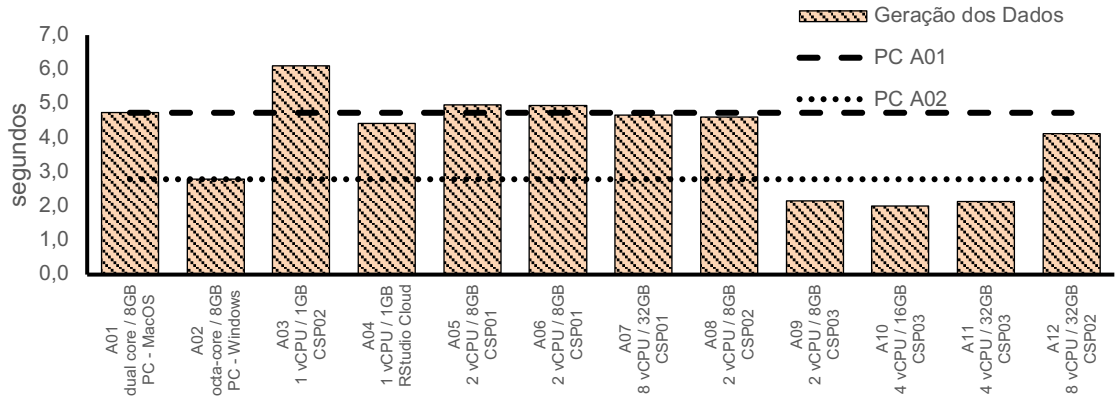

Figura 1. Média das execuções da etapa Geração de Dados do teste T1 (regressão simples em amostra normal)
Fonte: Resultados originais da pesquisa

Da mesma forma, as médias das execuções da etapa Execução do Modelo podem ser visualizadas na Figura 2, permitindo as observações:

- com exceção dos ambientes A01 (PC) e A03 (menor configuração em um dos provedores de nuvem – 1 CPU e 1GB de memória), todos os outros ambientes que concluíram o processamento desta etapa tem tempo de processamento similar ao do ambiente A02 (PC) e entre si, independente da configuração ou do provedor de nuvem.
- Os ambientes de nuvem A09, A10 e A11, como na etapa 1, também tiveram os melhores tempos de processamento.

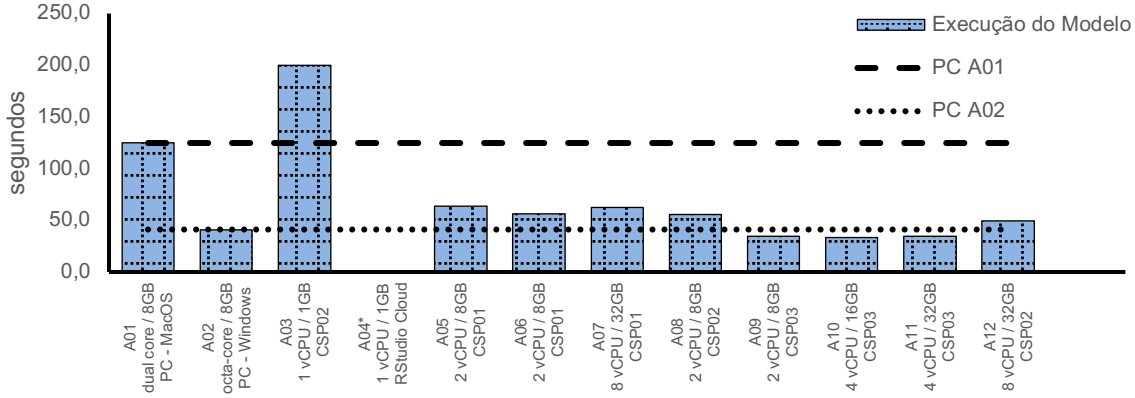

Figura 2. Média das execuções da etapa Execução do Modelo do teste T1 (regressão simples em amostra normal)
Fonte: Resultados originais da pesquisa
Nota: *Sem tempos - a etapa Execução do Modelo falhou para este ambiente

Cenário de teste T2 – Análise Exploratória em microdados do ENEM 2020

A Tabela 6 apresenta os resultados obtidos nas execuções do cenário de teste T2: Análise Exploratória nos microdados do ENEM 2020. Neste cenário foi excluído o ambiente A04 (RStudio, SaaS) que falhou no primeiro teste. Pode-se observar que as execuções nos ambientes A03, A05, A06 e A09, todos na nuvem mas com configurações diferentes, falharam na etapa de Inicialização / Geração dos Dados (leitura das bibliotecas, e carregamento do arquivo .csv de dados do ENEM.

Tabela 6. Resultados das execuções do teste T2 nos ambientes definidos para estudo

Ambiente	Execução	Inicialização[(1)]	Tratamento dos dados[(1)]	Execução do modelo[(1)]	Tempo total[(1)]
A01	#1	615,173	27,676	49,322	692,171
A01	#2	408,901	35,899	54,688	499,488
A01	#3	479,056	43,653	50,886	573,595
A02	#1	666,75	42,7	30,4	739,85
A02	#2	789,05	56,67	65,23	910,95
A02	#3	720,45	58,4	82,73	861,58
A03	#1	falha / erro	-	-	-
A03	#2	falha / erro	-	-	-
A03	#3	falha / erro	-	-	-
A05	#1	falha / erro	-	-	-
A05	#2	falha / erro	-	-	-
A05	#3	falha / erro	-	-	-
A06	#1	falha / erro	-	-	-
A06	#2	falha / erro	-	-	-
A06	#3	falha / erro	-	-	-
A07	#1	92,873	45,271	29,309	167,453
A07	#2	140,054	50,648	31,775	222,477
A07	#3	74,394	54,798	31,633	160,825
A08	#1	94,869	49,168	110,445	254,482
A08	#2	668,774	54,934	29,618	753,326
A08	#3	669,258	55,226	34,624	759,108
A09	#1	falha / erro	-	-	-
A09	#2	falha / erro	-	-	-
A09	#3	falha / erro	-	-	-
A10	#1	68,747	27,301	21,382	117,43
A10	#2	50,427	23,585	20,843	94,855
A10	#3	49,095	23,309	21,207	93,611
A11	#1	69,595	28,37	22,782	120,747
A11	#2	50,603	28,205	18,646	97,454
A11	#3	68,731	27,761	22,408	118,9
A12	#1	114,304	51,042	30,676	196,022
A12	#2	103,926	49,435	28,573	181,934
A12	#3	89,526	47,605	30,709	167,84

Fonte: Resultados originais da pesquisa
Nota: [(1)] Valores em segundos

A falha nos ambientes A03 e A04 ocorreu na carga dos dados, gerando uma mensagem vaga de erro após reiniciar o sistema automaticamente.

Por outro lado, nos ambientes A05 e A06, ambos executados no mesmo CSP, sendo a diferença apenas no tamanho do disco interno da VM provisionada, a falha ocorreu aproximadamente após uma hora de execução, durante o carregamento dos dados. Não houve mensagem de erro. Mas como estes são ambientes IaaS, foi possível acessar a console das VMs e constatar que os processos do RStudio não estavam mais ativos na VM. Nenhuma mensagem de erro foi gerada ou encontrada. Entretanto, observando-se os dados de monitoração disponíveis na console do ambiente de nuvem, e documentado para um dos testes na Figura 3, foi possível verificar que a carga dos dados no RStudio consome muita memória. Quando a memória física se exauri, a VM começa a fazer "swapping"[6] usando mais processador e disco. Esse fato pode ser observado em torno do tempo 15:20, nos gráficos "Uso da memória", "Uso da CPU" e "IOPS[7] do disco".

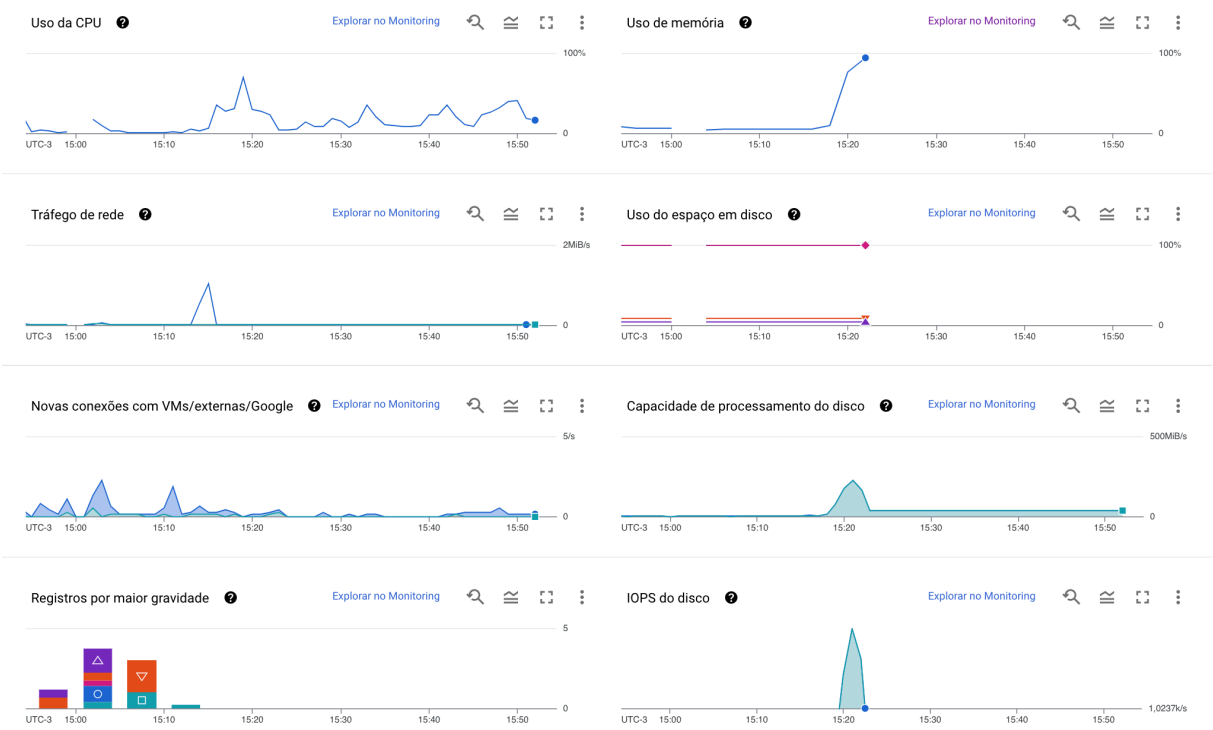

Figura 3. Métricas da VM durante a execução do teste T2 em ambiente de nuvem
Fonte: Resultados originais da pesquisa – serviço de "Google Cloud Monitoring" para a VM GCP em execução

[6] Quando os dados de um programa em execução não cabem na memória principal ("Random Access Memory [RAM]") instalada no equipamento, o sistema operacional usa a memória secundária (chamada de área de "swap") para armazená-los. Como a CPU só acessa a RAM diretamente, durante o processamento pode ser necessário mover outros dados armazenados na RAM para a área de "swap", e depois trazê-los de volta para a RAM. Essas operações de troca de dados entre os níveis de memória são chamadas de "swapping" ou operações de "swap", e torna o processamento dos dados mais lento do que se fosse usada somente a memória RAM (Darú, 2018).
[7] "Input/Output Operations per Second [IOPS]"

A Figura 4 traz a representação gráfica da média dos tempos de execução de cada etapa em cada ambiente, comparando os tempos dos PCs com os das execuções na nuvem. Considerações importantes podem ser obtidas desta visualização:

- Os ambientes A07, A10, A11 e A12 são ambientes na nuvem e são também os que possuem maior memória: 32GB, 16GB, 32GB e 32GB respectivamente. Como a fase Inicialização é a que requer mais memória, o fato de estes ambientes terem mais memória permitiu o menor tempo de processamento no carregamento dos dados. Ainda neste raciocínio pode-se notar, pelo resultado mais rápido do ambiente A10 com 16GB de memória, que após atingir a capacidade necessária para processamento, aumentar a memória pode não trazer nenhum benefício para tempo de processamento.
- Já os ambientes A10 e A11 tiveram um tempo de processamento menor nas etapas Tratamento dos Dados e Execução do Modelo, mesmo não tendo os maiores processadores dos ambientes testados, ou a última versão do R.
- Considerando-se a fase Execução do Modelo, nota-se que usar um ambiente com o dobro de CPU (A10 e A11) diminui para aproximadamente a metade o tempo de execução (comparando-se o ambiente A01, PC que foi usado para codificação e testes iniciais, com os ambientes com melhor performance).

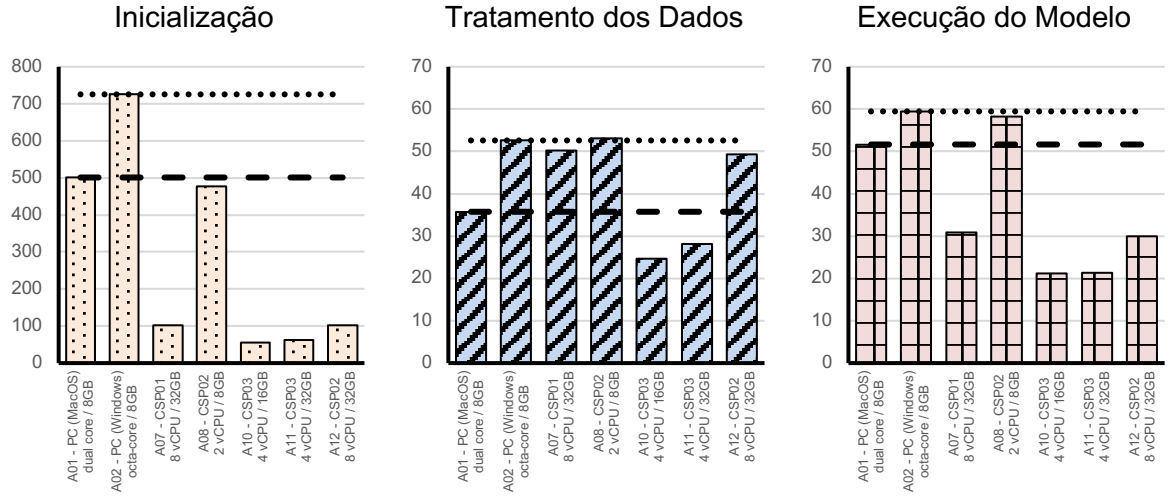

Figura 4. Média das execuções das etapas do teste T2 (Análise Exploratória em microdados do ENEM 2020)
Fonte: Resultados originais da pesquisa
Nota: Tempos em segundos

Levando-se em consideração que para executar um modelo é necessário preparar os dados (limpar, tratar e validar), como explicado por Patil e Hiremath (2018), e para isso é necessário que os dados sejam carregados no ambiente para análise, é importante não

apenas considerar os tempos individuais de cada etapa (que são úteis para o entendimento do uso dos recursos), mas também observar o tempo total que reflete o uso completo de memória e CPU requerido. A Figura 5 traz a representação gráfica da média dos tempos totais da execução do modelo T2 – Análise Exploratória dos microdados do ENEM 2020, que permite a observação de algumas tendências:

- Ambientes baseados em Unix (MacOS – A01) ou Linux (Ubuntu – A08), com configurações semelhantes, provisionado na nuvem ou não, tem tempo de processamento semelhantes.
- Ambiente Windows (PC – A02) tem o maior tempo total de processamento.
- Ambientes em nuvem, provisionados com maior capacidade de memória e CPU (A07, A10, A11 e A12) tem o melhor processamento. Entretanto, o ambiente que tem o melhor tempo de processamento é o A10, que dentre estes quatro é o que tem menor capacidade: 4 vCPU e 16GB e Memória – contra 8 vCPU e 32 GB de memória dos ambientes A08 e A13 e 4 vCPU e 16 GB de memória.

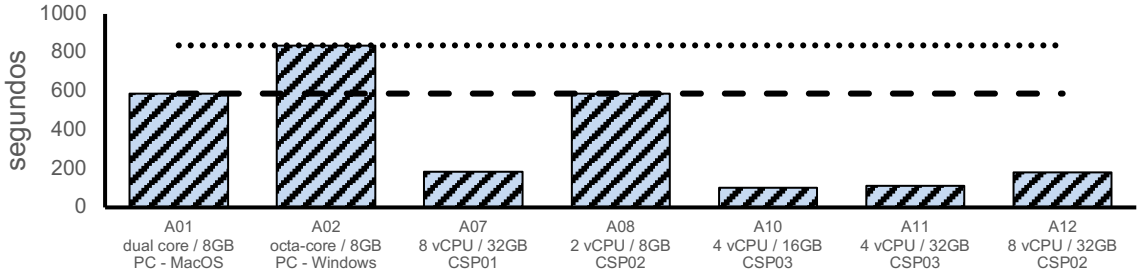

Figura 5. Média dos tempos totais das execuções do teste T2 (Análise Exploratória em microdados do ENEM 2020)
Fonte: Resultados originais da pesquisa

Cenários de teste T3 e T3r – Análise de Agrupamentos e Análise de Correspondência Múltipla [ACM] dos microdados do ENEM 2020

Inicialmente, tendo em vista o tamanho do dataset de estudo, um sub-dataset foi criado com dados de apenas um Estado brasileiro (Roraima), e esse sub-dataset tornou-se o teste T3 reduzido [T3r].

Considerando que os ambientes não completaram o teste T2 (A03, A05, A05 e A09), falharam na etapa Inicialização que é comum com este teste T3, estes ambientes não foram

considerados para este cenário. Da mesma forma o ambiente A04, que foi removido dos testes após falhar no teste T1.

Os resultados das execuções dos testes T3r podem ser observados na Tabela 7. Nela nota-se que a execução para o ambiente A08 falhou na etapa Análise de Agrupamentos. A falha ocorreu ao tentar executar a função de identificação da quantidade de agrupamentos ideal para o modelo, com o método “Gap statistic method”.

Tabela 7. Resultados das execuções do teste T3r nos ambientes definidos para estudo

Ambiente	Execução	Inicialização(1)	Tratamento dos dados(1)	Análise de Agrupamentos(1)	ACM(1)	Tempo Total(1)
A01	#1	750,175	4,170	3392,073	10,600	4157,018
A01	#2	519,557	12,676	3456,340	6,137	3994,710
A01	#3	619,009	9,697	4947,066	5,046	5580,818
A02	#1	685,200	9,090	2869,420	15,200	3578,910
A02	#2	679,170	11,160	2719,250	5,890	3415,470
A02	#3	815,430	10,920	2967,750	7,810	3801,910
A07	#1	70,545	0,330	2889,353	3,552	2963,780
A07	#2	81,134	26,040	2837,975	3,708	2948,857
A07	#3	73,866	0,405	2825,088	3,775	2903,134
A08	#1	533,170	57,288	falha / erro	-	-
A08	#2	660,523	49,447	falha / erro	-	-
A08	#3	729,704	68,695	falha / erro	-	-
A10	#1	63,908	0,769	998,388	2,881	1065,946
A10	#2	65,439	2,308	981,636	3,872	1053,255
A10	#3	45,077	0,171	937,021	2,535	984,804
A11	#1	51,922	0,194	1067,233	3,083	1122,432
A11	#2	42,974	0,175	1037,447	1,781	1082,377
A11	#3	44,204	0,187	1015,409	3,335	1063,135
A12	#1	71,193	0,378	2954,116	2,479	3028,166
A12	#2	85,113	12,370	2919,844	3,648	3020,975
A12	#3	72,003	10,896	2869,080	3,548	2955,527

Fonte: Resultados originais da pesquisa
Nota: (1) Valores em segundos

A Figura 6, que mostra um comparativo visual da média das execuções do teste em cada ambiente, permite observar que o processamento nos ambientes de nuvem (A07, A10, A11 e A12) tiveram menor tempo de processamento. No caso de A10 e A11, a melhora foi de aproximadamente 80%. Nesta observação é importante também notar que estes dois ambientes não são os de maior CPU, tendo apenas 4 CPUs contra 8 CPUs dos ambientes A07 e A12. A diferença entre A10 e A11, de melhor tempo de processamento e A07 e A12 é o provedor de nuvem.

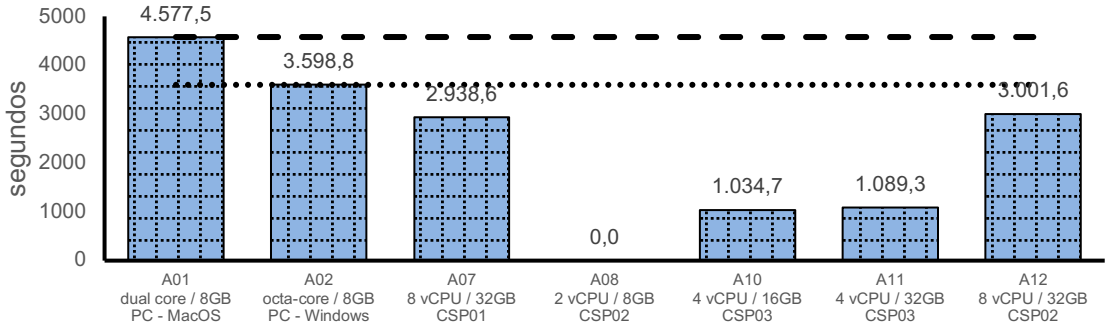

Figura 6. Média dos tempos totais das execuções do teste T3 reduzido (Análise de Agrupamentos e ACM dos microdados do ENEM 2020 para o Estado de Roraima)
Fonte: Resultados originais da pesquisa

Já o teste T3 com todos os microdados do ENEM 2020 falhou em todos os ambientes durante o cálculo da matriz de distância pelo método “euclidean”. A mensagem de erro, “Error: cannot allocate vector of size 23299,9 Gb”, indica que a função necessitava mais memória para concluir do que a disponível, mesmo nos ambientes maiores.

Cenário de teste T4 – Análise Fatorial por Componentes Principais [PCA] dos microdados do ENEM 2020

Os resultados obtidos no teste T4, PCA, estão documentados na Tabela 8. Pode-se observar nesta tabela que todos os ambientes testados concluíram a execução dos “scripts”. Ambientes não testados foram os que falharam no teste T1, e os que falharam em etapas comuns (Inicialização ou Tratamento dos Dados) nos testes T2 e T3/T3r.

Tabela 8. Resultados das execuções do teste T4 nos ambientes definidos para estudo
(continua)

Ambiente	Execução	Inicialização(1)	Tratamento dos dados(1)	PCA(1)	Tempo Total(1)
A01	#1	240,803	46,783	1042,069	1329,655
A01	#2	319,300	47,847	942,223	1309,370
A01	#3	761,477	57,020	930,926	1749,423
A02	#1	847,100	92,140	744,910	1684,150
A02	#2	763,240	53,220	695,710	1512,170
A02	#3	914,830	115,360	723,530	1753,720
A07	#1	72,522	44,730	914,239	1031,491
A07	#2	75,910	51,117	951,124	1078,151
A07	#3	105,620	58,789	961,999	1126,408
A08	#1	666,904	76,494	576,764	1320,162
A08	#2	531,084	139,984	603,271	1274,339
A08	#3	755,254	109,981	568,796	1434,031

Tabela 8. Resultados das execuções do teste T4 nos ambientes definidos para estudo (conclusão)

Ambiente	Execução	Inicialização[(1)]	Tratamento dos dados[(1)]	PCA[(1)]	Tempo Total[(1)]
A10	#1	42,822	26,614	313,631	383,067
A10	#2	42,366	25,777	313,591	381,734
A10	#3	42,326	57,505	315,971	415,802
A11	#1	44,553	29,579	316,561	390,693
A11	#2	43,172	28,607	311,719	383,498
A11	#3	42,137	25,346	313,999	381,482
A12	#1	79,985	56,333	624,028	760,346
A12	#2	71,360	50,355	592,084	713,799
A12	#3	73,212	48,672	602,387	724,271

Fonte: Resultados originais da pesquisa
Nota: [(1)] Valores em segundos

Estes dados mostrados na Figura 7, em um gráfico das médias dos tempos de execução por ambiente, permitem a visualização e comparação dos diferentes ambientes e as seguintes observações:

- Nota-se que as execuções em ambientes de nuvem A07, A10, A11 e A12 tiveram um tempo de processamento melhor que os dois PCs (A01 e A02).
- Como a diferença das médias de tempos entre A02/A01 e A08/A01 foi menor que 10% e considerando a quantidade de execuções do teste em cada ambiente, pode-se considerar que os três ambientes, de mesma configuração (memória e CPU), possuem tempos de processamento semelhantes.
- Os ambientes em nuvem com mais memória e CPU, A07, A10, A11 e A12, têm tempos de processamento menores, quando comparados com o ambiente em nuvem A08, com menor capacidade.
- Pela comparação de A07 (8 CPUs / 32GB memória), A10 (4 CPUs / 16GB memória), A11 (4 CPUs / 32GB memória) e A12 (8 CPUs / 32GB memória), nota-se que não foram as máquinas com maior capacidade de processamento (CPU) que tiveram melhores resultados. Essa observação permitiu o entendimento que outros critérios, não considerados neste estudo, também influenciam no tempo de processamento e que aumentar a capacidade além da necessária não garante melhora do tempo de processamento.

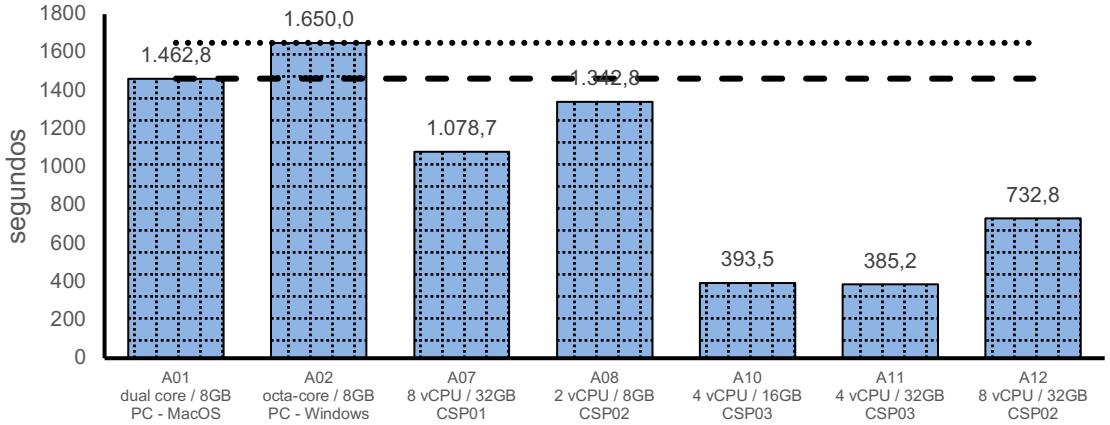

Figura 7. Média dos tempos totais das execuções do teste T4
Fonte: Resultados originais da pesquisa

Revisão dos custos de cada cenário de teste

Antes de avaliar os custos, cabe a verificação dos tempos totais da realização dos testes em cada ambientes.

A Tabela 9 permite a visualização completa dos testes, contendo o tempo de processamento médio em cada cenário em segundos e a totalização por ambiente. Foram incluídos também os ambientes e testes que falharam em alguns ou todos os cenários para permitir a análise do panorama completo. Nota-se que o ambiente com maior tempo médio de processamento foi o PC A01 (01h53min), seguido de um ambiente de nuvem com configuração similar, 2 CPUs e 8GB de memória, A07 (01h11min); e o ambiente com melhor, ou seja, menor tempo de processamento, não foi o ambiente de nuvem com mais capacidade, mas o ambiente com 4 CPUs e 16GB de memória, A10 (00h26min).

Tabela 9. Resultados das execuções dos testes e tempo total de processamento por ambiente

Ambiente	Tempo de Processamento médio				Tempo total		Tempo de
	T1	T2	T3r	T4	segundos	horas	processamento
A01	129,49	588,42	4577,52	1462,82	6758,23	1,88	01h53min
A02	43,86	837,46	522,85	1650,01	3054,19	0,85	00h51min
A03	242,63	-	-	-	242,63	0,07	00h04min
A04	-	-	-	-	-	-	-
A05	68,35	-	-	-	68,35	0,02	00h02min
A06	62,67	-	-	-	62,67	0,02	00h02min
A07	67,08	183,59	2938,59	1078,68	4267,94	1,19	01h11min
A08	60,55	588,97	-	1342,84	1992,36	0,55	00h33min
A09	36,92	-	-	-	36,92	0,01	00h01min
A10	35,62	101,97	1034,67	393,53	1565,79	0,43	00h26min
A11	36,72	112,37	1089,31	385,22	1623,62	0,45	00h27min
A12	53,59	181,93	3001,56	742,49	3979,57	1,11	01h07min

Fonte: Resultados originais da pesquisa

Com base nos tempos de processamento de cada ambiente, a avaliação do custo relacionado a eles pôde ser calculada. A Tabela 10 apresenta, para os ambientes que completaram todos os testes, as informações necessárias para cálculo do custo total, tanto em dólares americanos das configurações dos ambientes nos CSP, como em reais brasileiros, convertidos a partir da taxa de câmbio comercial para compra de 30 set. 2022 (R$ / US$: 5,406), disponível no site do Instituto de Pesquisa Econômica Aplicada [IPEA][8] do Governo Brasileiro. Vale ressaltar nesta análise, que os PCs A01 e A02 refletem custo de aquisição do equipamento, enquanto nos ambientes na nuvem paga-se somente a utilização dos recursos.

Tabela 10. Cálculo do custo total de cada cenário por ambiente testado

Ambiente	tipo	Tempo de processamento	Custo (US$)		Custo total (R$)
			por hora	total	
A01	PC	02h00min	-	-	15300,00
A02	PC	01h00min	-	-	5800,00
A07	CSP	02h00min	0,28	0,56	3,03
A10	CSP	01h00min	0,17	0,17	0,92
A11	CSP	01h00min	0,23	0,23	1,24
A12	CSP	02h00min	0,23	0,46	2,49

Fonte: Resultados originais da pesquisa

Exemplos de custos não considerados para os ambientes A01 e A02 (PCs / “On-Premise”), mas que podem afetar o custo final: eletricidade, manutenção do equipamento, “firewall”, “upgrade” ou reposição do equipamento ou um “nobreak” para evitar indisponibilidades durante o processamento, dentre outros. Da mesma forma, para os ambientes em nuvem há custos não considerados na análise, como elementos de rede (egresso de dados do ambiente de nuvem), armazenamento (“storage”) ou mesmo o equipamento para acessar esses recursos de forma remota – que neste caso pode ser um PC qualquer com capacidade de processamento o suficiente para permitir o uso de um navegador.

Um PC pode ter outros usos, em momentos separados ou ao mesmo tempo. Em momentos separados, toda a capacidade de processamento do equipamento seria dedicada ao processamento da análise. Em contrapartida, outros usos em paralelo ao processamento da análise dividiria a capacidade do equipamento, impactando uma ou ambas as atividades.

Enquanto para o uso de um equipamento pessoal ou “on-premise”, com as mesmas configurações que um equivalente em nuvem, o profissional terá maiores despesas de capital (“Capital Expediture [CAPEX]”), precisando investir ou desembolsar previamente recursos financeiros para aquisição do equipamento, na nuvem o investimento maior será em despesas

[8] IPEA: http://www.ipeadata.gov.br/ExibeSerie.aspx?serid=38590&module=M

operacionais (“Operational Expediture [OPEX]”): o profissional paga pelo que usar. Importante também considerar os custos de equipamento para acesso à nuvem, mas estes não precisam ser equipamentos com maior capacidade (CPU e memória), o que torna o custo deles menor. Assim, não há OPEX neste cenário sem um valor mínimo de CAPEX.

Ainda, se o mesmo profissional investir em um determinado equipamento com específica configuração necessária para trabalhar sua análise e em algum momento o projeto precisar de um equipamento melhor, ele precisará investir novamente gerando um CAPEX maior para aquele projeto. Em contrapartida, no ambiente em nuvem, dada a mesma situação, o profissional poderá fazer um upgrade ou provisionar novo ambiente com as novas configurações necessárias e migrando a imagem da máquina original. Não há a necessidade de pagar pelo equipamento, apenas pelo que usou da configuração original e, ao final, pelo que usar da nova configuração provisionada.

Comparação de custos para estabelecimento de critérios para seleção

Considerando que um dos critérios para seleção do CSP selecionados foi performance, este estudo comparou o custo do PC / “on-premise” com os ambientes em nuvem que tiveram tempo total de execução de todos os scripts menor ou igual a uma hora.

O custo do ambiente “on-premise” / PC é CAPEX. Não considerando os custos de upgrade e manutenção, esse custo será único, ou seja, não terá variação ao longo dos meses.

Por outro lado, o custo dos ambientes em nuvem é OPEX e depende da utilização, aumentando assim com o tempo.

A Tabela 11 apresenta uma projeção dos custos acumulados ao longo dos meses para os ambientes com melhor desempenho, ou seja, tempo total de execução de todos os scripts deste estudo igual ou menor a uma hora – A02 (PC) e A10 e A11 (nuvem), incluindo:

- o custo do ambiente “On-Premise” / PC que é constante ao longo do tempo – o investimento é feito somente uma vez;
- o custo dos ambientes em nuvem que aumenta durante o tempo, e neste caso, considerou-se utilização contínua 24h/dia durante 30 dias por mês.

Tabela 11. Projeção de custo acumulado dos ambientes com tempo total de execução menor ou igual a 1h

ambiente	custo / h	1 mês	2 meses	3 meses	4 meses	5 meses	6 meses	7 meses	8 meses	9 meses
A02	5800,00	5800,00	5800,00	5800,00	5800,00	5800,00	5800,00	5800,00	5800,00	5800,00
A10	0,92	662,40	1324,80	1987,20	2649,60	3312,00	3974,40	4636,80	5299,20	5961,60
A11	1,24	892,80	1785,60	2678,40	3571,20	4464,00	5356,80	6249,60	7142,40	8035,20

Fonte: Resultados originais da pesquisa
Nota: Valores em Reais [R$]

A Figura 8 traz a representação visual da evolução do custo para os ambientes com melhor tempo de processamento, permitindo uma melhor análise da evolução dos custos. Nela pode-se observar que o custo acumulado de A10 (nuvem) torna-se maior que o custo do PC / “On-Premise” perto do 9º mês de utilização contínua; e o ambiente A11 (nuvem), entre o 6º e 7º mês. Através deste gráfico, é possível avaliar o melhor cenário de custo para execução dos cenários propostos, levando-se em conta o tempo de retorno do projeto ou as expectativas de investimento.

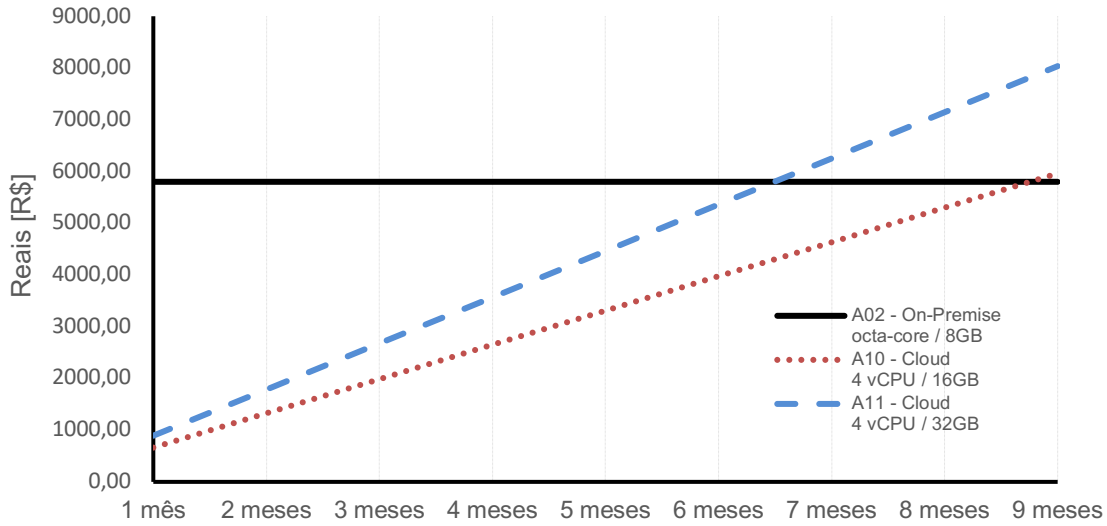

Figura 8. Projeção de Custos acumulados dos ambientes com melhor performance
Fonte: Resultados originais da pesquisa

Gangadhar e Shaikh (2021) mencionam que o cálculo do Retorno Sobre Investimento (“Return Of Investment [ROI]”) é uma das partes mais desafiadoras da adoção do ambiente em nuvem e inclui vários elementos de custo, que muitas vezes são ignorados, além de um período definido para a análise financeira dos investimentos. Assim, é importante ressaltar que somente a análise deste estudo não é suficiente para calcular o ROI completo do uso de ambiente em nuvem em comparação com um ambiente “On-Premise” / PC. Antes, este estudo trouxe considerações para avaliação dos ambientes frente aos critérios de performance e custo relacionado adotados.

Considerações Finais

A escolha do melhor ambiente para trabalhar dados depende de necessidades e avaliações individuais: um ambiente que é o melhor para alguém ou um projeto pode não ser a melhor escolha para outro.

Ter um equipamento pessoal ou “On-Premise” para trabalhos de análise de dados, além de garantir um custo conhecido e fixo, permite também outros usos – e mesmo, o acesso aos recursos em nuvem.

O ambiente na nuvem não é mais ou menos performático que um equipamento pessoal: depende da configuração escolhida. Mas as vantagens de um ambiente na nuvem estão em um menor investimento financeiro inicial, no fato de poder migrar de máquina (ou fazer “upgrade” da máquina original) quando mais capacidade de processamento é necessária para a análise de dados, e no fato de o custo envolvido ser somente do tempo de uso.

Quanto ao custo, inicialmente as opções na nuvem parecem mais interessantes. Mas há ainda que se considerar o uso – tempo e recursos -, e a administração dos gastos: tanto a monitoração do uso e custos, quanto a parada dos recursos alocados para evitar cobranças sem uso efetivo. Importante notar também que a escolha do ambiente correto na nuvem é um critério relevante para evitar custos desnecessários com recursos não utilizados por conta de uma configuração de capacidade muito acima da necessária.

Ainda uma vantagem a se considerar no uso do ambiente de nuvem é, para pequenas empresas e instituições de ensino, a padronização dos equipamentos para os profissionais e alunos. Com todos profissionais ou professores e alunos usando o mesmo ambiente, a replicação dos testes ou a análise de erros e problemas torna-se muito mais simples para a equipe de suporte.

Com o constante e certo aumento de dados disponíveis, a análise de dados torna-se mais que uma mera curiosidade, mas uma necessidade do ser humano e das empresas para melhores decisões e crescimento. Computação em nuvem não é mais algo restrito somente a grandes empresas, mas é algo acessível a todos: profissionais independentes e pequenas e médias empresas. Essa acessibilidade permite mais exploração e entendimento dos dados. Algo que antes era restrito a grandes empresas, capazes de absorver o custo dos ambientes “On-Premise”, é agora aberto a todos, com opções de investimento iniciais maiores (PC / “On-Premise”) ou distribuído ao longo dos meses (nuvem).

Este estudo não tinha o propósito de avaliar todos os fatores que impactam a análise de dados, em nuvem ou “On Premise”, uma vez que limitou seu foco em um determinado modelo de serviço de nuvem e em alguns parâmetros de análise (algoritmos e dados). Seu objetivo foi apresentar alguns critérios para adoção do melhor ambiente para análise, dadas as necessidades individuais de cada pessoa ou empresa. Assim, há inúmeras possibilidades para extensão deste estudo em trabalhos futuros, com outros algoritmos de aprendizado de máquina, ou os mesmos com outros conjuntos de dados, exploração de plataformas específicas para “machine learning” e ciências de dados criadas pelos provedores de nuvem,

ou o uso de banco de dados como recurso para armazenamento e distribuição da carga de processamento, ou ainda, a quebra do “dataset” em partições menores para processamento paralelo, tanto em um PC quanto na nuvem, e por fim, as próprias inovações em termos de recursos e serviços que são parte das ofertas de nuvem.

Agradecimentos

Ao meu marido Ricardo, pelo suporte, paciência e conselhos ao me ouvir falar (muitas vezes) sobre o tema; ao meu filho Marcos Vinícius pelo carinho e momentos de descontração; à minha mãe Dalney, que de forma indireta, me deu a ideia da análise para usar nos testes deste estudo; e aos meus colegas profissionais, pelas discussões e suporte para este trabalho.

Referências

Alabool, H.; Kamil, A.; Arshad, N.; Alarabiat, D. 2018. Cloud service evaluation method-based Multi-Criteria Decision-Making: A systematic literature review. In: ScienceDirect, Journal of Systems and Software, Volume 139: 161-188. Disponível em https://www.sciencedirect.com/science/article/pii/S0164121218300244. Acesso em 29 jun. 2022.

Darú, C.D.H. 2018. Comparação de desempenho de caches de segmentação e de paginação. Universidade Federal do Paraná, Curitiba, PR, Brasil. Disponível em https://www.inf.ufpr.br/roberto/tgClara.pdf. Acesso em 30 set. 2022.

Fávero, L.P.; Belfiori, P. 2017. Manual de análise de dados. 1ed. Elsevier, Rio de Janeiro, RJ, Brasil.

Gangadhar, V.R.; Shaikn, A. 2021. Cloud Technology and Return of Investiment (ROI). In Special Issue of First International Conference on Engineering, Science and Technology (ICEST 2021), Research Journal on Advanced Science Hub (IRJASH), Volume 03 Issue 01S January 2021. Disponível em https://rspsciencehub.com/article_8086.html. Acesso em 03 out. 2022.

Gerhardt, T.E.; Silveira, D.T. 2009. Métodos de Pesquisa. Editora da Universidade Federal do Rio Grande do Sul [UFRGS], Porto Alegre, RS, Brasil. Disponível em: https://www.lume.ufrgs.br/bitstream/handle/10183/213838/000728731.pdf. Acesso em 09 jul. 2022.

Herbst, N.R.; Kounev, S.; Reussner, R. 2013. Elasticity in Cloud Computing: What It Is, and What It Is Not. In USENIX Association, 10th International Conference on Autonomic Computing (ICAC’13). Disponível em https://www.usenix.org/conference/icac13/technical-sessions/presentation/herbst. Acesso em 30 set. 2022.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP]. 2020. ENEM, Microdados, Dados Abertos. Disponívei em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem. Acesso em 09 jul. 2022.

Kandula, S.; Li, A.; Yang X.; Zhang, M. 2010. CloudCmp: comparing public cloud providers. In: Proceedings of the 10th ACM SIGCOMM conference on Internet measurement (IMC '10). Association for Computing Machinery, New York, NY, USA, 1–14. Disponível em https://dl.acm.org/doi/abs/10.1145/1879141.1879143. Acesso em 07 jul. 2022

Kumari, A.; Verma, N. 2019. An Analytical Review Study on Big Data Analysis using R Studio. In: International Journal of Engineering Technologies and Management Research, Volume 6. Disponível em https://www.academia.edu/39738586/AN_ANALYTICAL_REVIEW_STUDY_ON_BIG_DATA_ANALYSIS_USING_R_STUDIO. Acesso em 05 ago. 2022.

Lane, M.; Shrestha, A.; Ali, O. 2017. Managing the Risks of Data Security and Privacy in the Cloud: A Shared Responsibility between the Cloud Service Provider and the Client Organisation. University of Southern Queensland, Queensland, Australia. Disponível em https://eprints.usq.edu.au/. Acesso em 30 set 2022.

Lang, M.; Wiesche, M.; Krcmar, H. 2018. Criteria for Selecting Cloud Service Providers: A Delphi Study of Quality-of-Service Attributes. In: ScienceDirect, Journal of Information & Management, Volume 55, Número 6: Pages 746-758. Disponível em https://www.sciencedirect.com/science/article/pii/S0378720617303142. Acesso em 29 jun. 2022.

Lynn, T.; Mooney, J.G.; 2020. Measuring the Business Value of Cloud Computing. eBook, Open Access. Palgrave Mcmillan, Cham, Suiça. Disponível em https://library.oapen.org/bitstream/handle/20.500.12657/41747/2020_Book_MeasuringTheBusinessValueOfClo.pdf?sequence=1#page=39.

National Institute of Standards and Technology [NIST]. 2011. The NIST Definition of Cloud Computing – Special Publication 800-145.
Disponível em: https://csrc.nist.gov/publications/detail/sp/800-145/final
https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf. Acesso em 09 jul. 2022.

Patil, M.M.; Hiremath, B.N. 2018. A Systematic Study of Data Wrangling. In: International Journal of Information Technology and Computer Science. Modern Education and Computer Science [MECS] press. Disponível em https://www.mecs-press.com/ijitcs/ijitcs-v10-n1/IJITCS-V10-N1-4.pdf. Acesso em 30 set. 2022.

Porto, F.; Ziviani, A. 2014. Ciência de dados. III Seminário de Grandes Desafios da Computação no Brasil, Rio de Janeiro, RJ. Disponível em:
https://www.lncc.br/~ziviani/papers/III-Desafios-SBC2014-CiD.pdf

Weinman, J. 2016. Hybrid Cloud Economics. In: IEEE Cloud Computing, Volume 3, número 1: 18-22. Disponível em https://ieeexplore.ieee.org/abstract/document/7420473. Acesso em 07 jul. 2022

World Economic Forum [WEF] – Vopson, Melvin M. 2021. The world's data explained: how much we're producing and where it's all stored. Disponível em:
https://www.weforum.org/agenda/2021/05/world-data-produced-stored-global-gb-tb-zb/.
Acesso em 07 abr. 2022.