MBA USP ESALQ

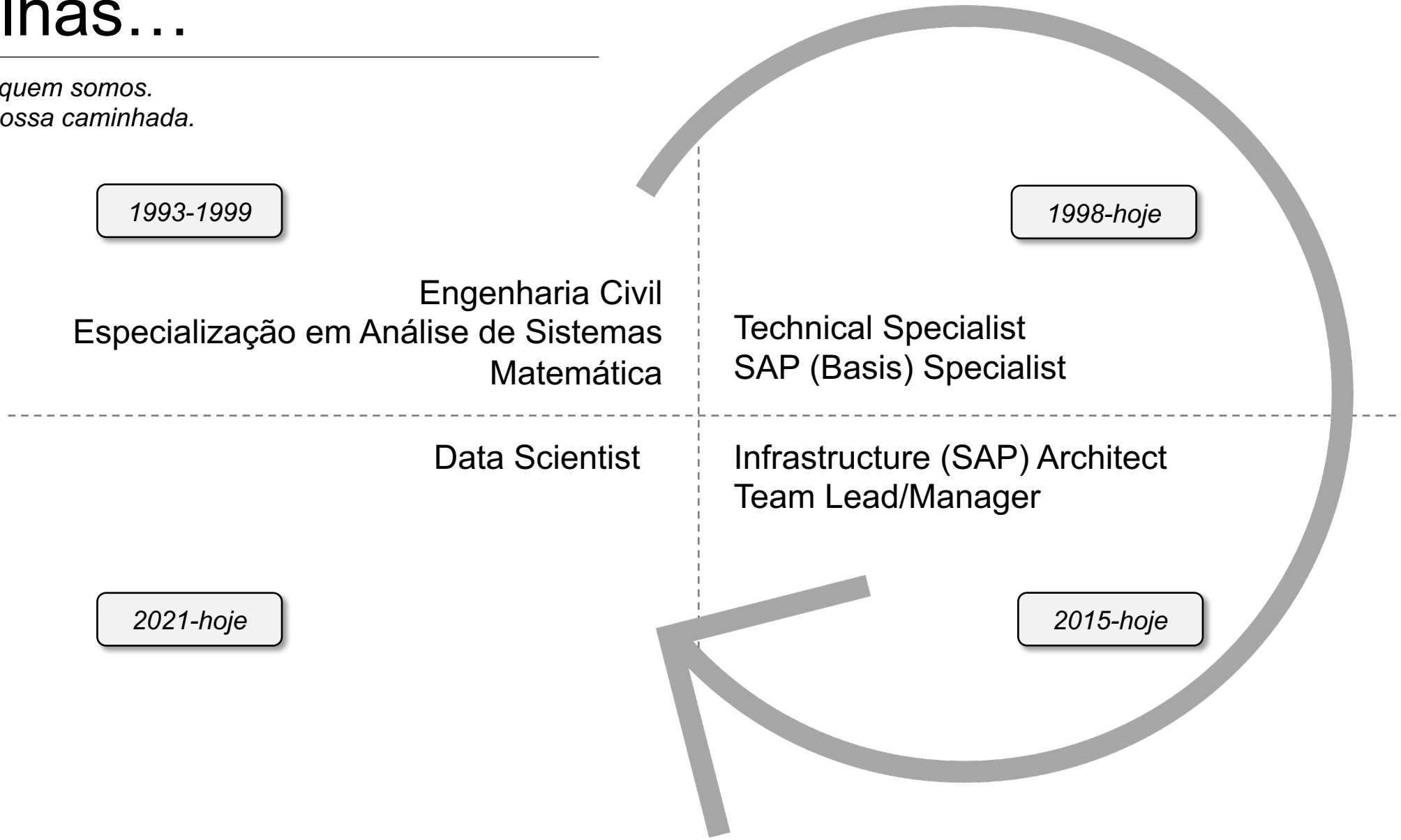
Critérios para adoção de computação em nuvem para execução de técnicas de aprendizado de máquina

Adriana Melges Quintanilha Weingart
Thiago Bianchi

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

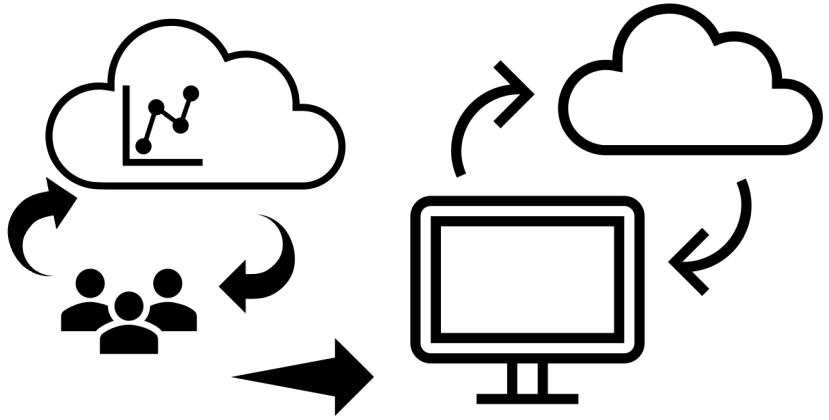
Escolhas…

… definem quem somos.
… afetam nossa caminhada.



Escolhas e experiências…

… basearam a definição do tema.



Registro de dados crescente – interesse e importância na análise destes dados.
Análise por pessoas, profissionais autônomos, pequenas / médias empresas.
Uso de Computação em Nuvem / Cloud Computing →

- É caro?
- Como escolher?
- É realmente mais performático?
- É opção?

As definições do trabalho…

… Materiais e métodos

Definições

- Dados e algoritmos de teste
  - Teste inicial
  - Dados do ENEM 2020
  - Análise Exploratória, PCA, Cluster e ACM

  Scripts disponíveis em https://bit.ly/AW_TCC_Scripts

- Provedores de serviço de “cloud”
  - Quadrantes do Gartner
  - Seleção de 3 Provedores de Cloud

- Ambientes de teste
  - Equipamentos pessoais (MAC & Windows)
  - VMs em Cloud (Linux) - variando capacidade

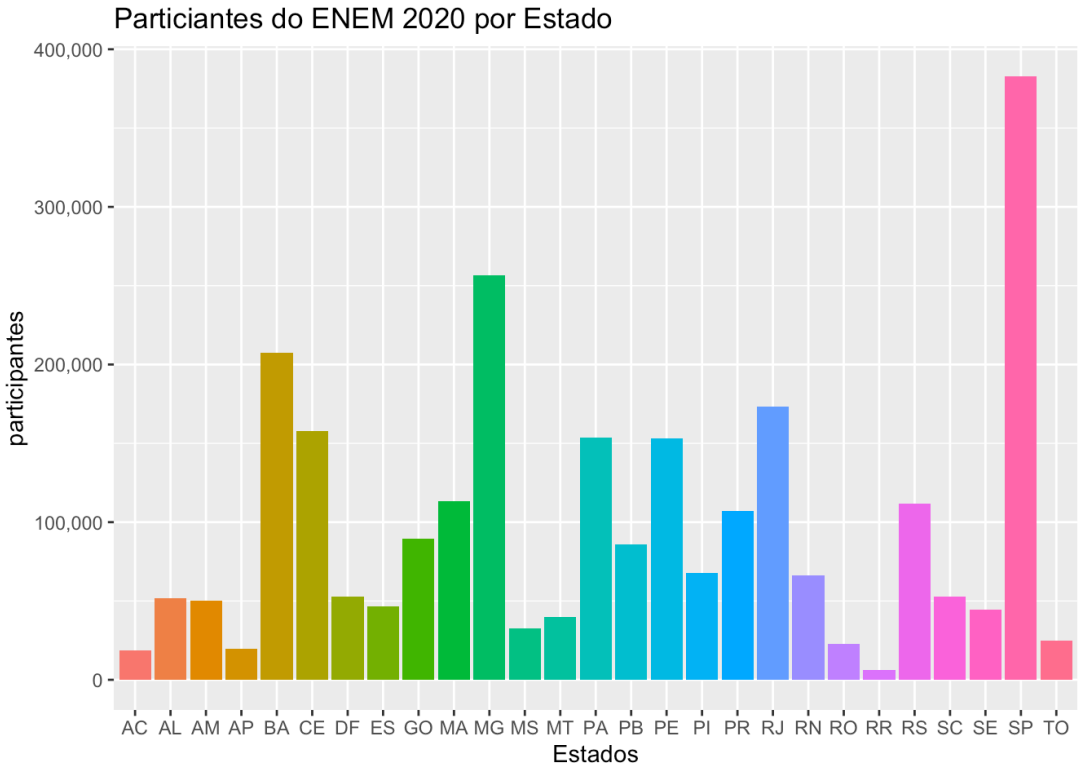
  Procedimentos disponíveis em https://bit.ly/AW_TCC_Procedimentos

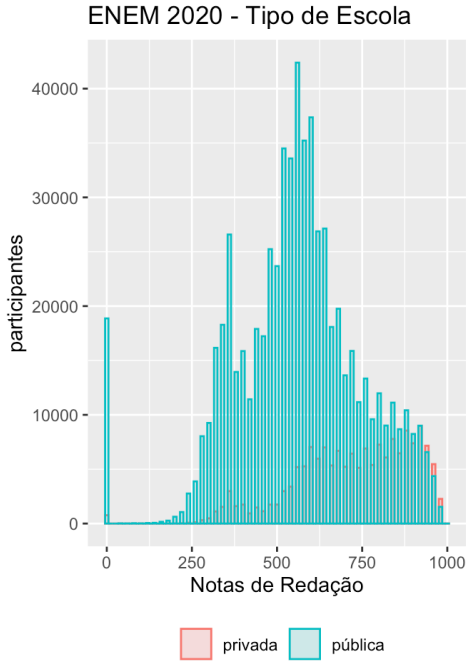
- Critérios para análise e comparação
  - Estudos anteriores identificando critérios
  - Seleção dos 3 principais critérios

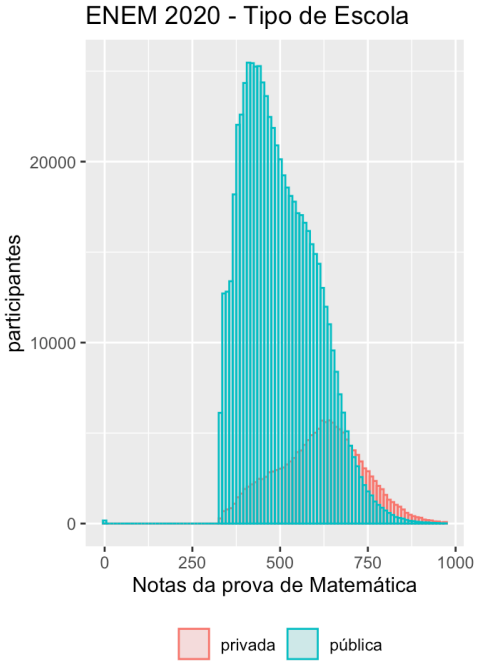
- Ferramentas de visualização
  - R / RStudio – para algoritmos de teste
  - Excel – para resultados do estudo
  - Coblis – para validação cores/gráficos

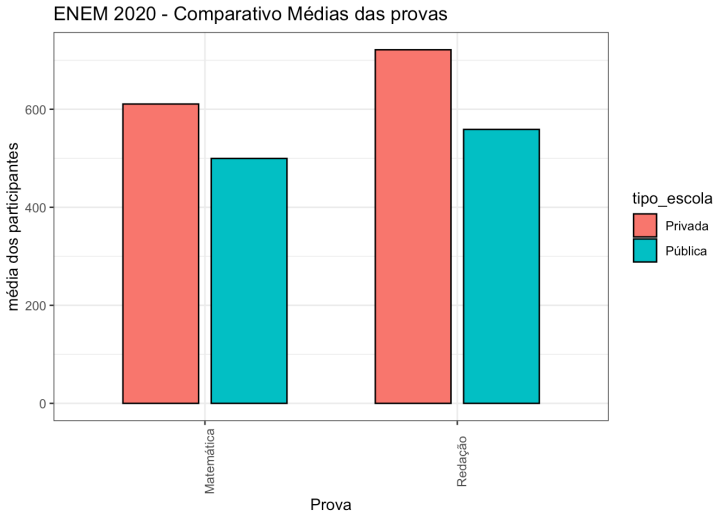
Relatórios gerados – Análise Exploratória

Materiais e métodos
Relatórios (R markdown) disponíveis em https://bit.ly/AW_TCC_ENEM-2020





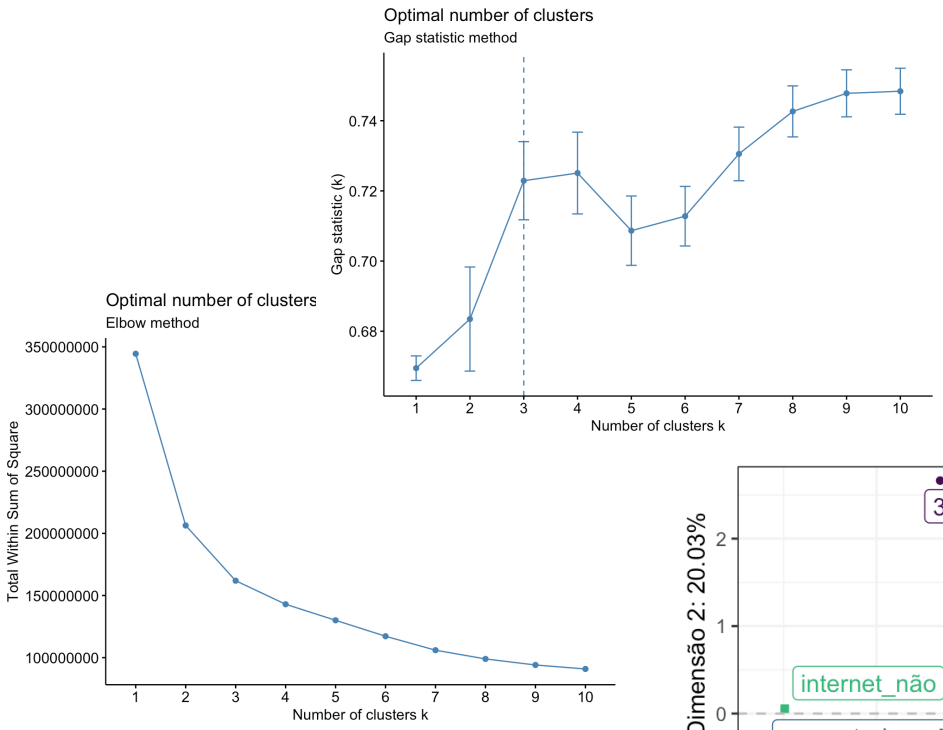




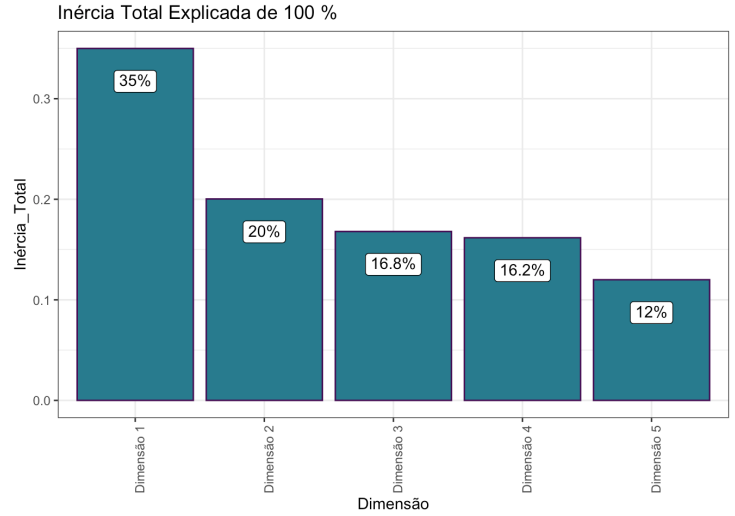
Análise exploratória para conhecimento dos dados, do volume de participantes por Estado, e comparações iniciais

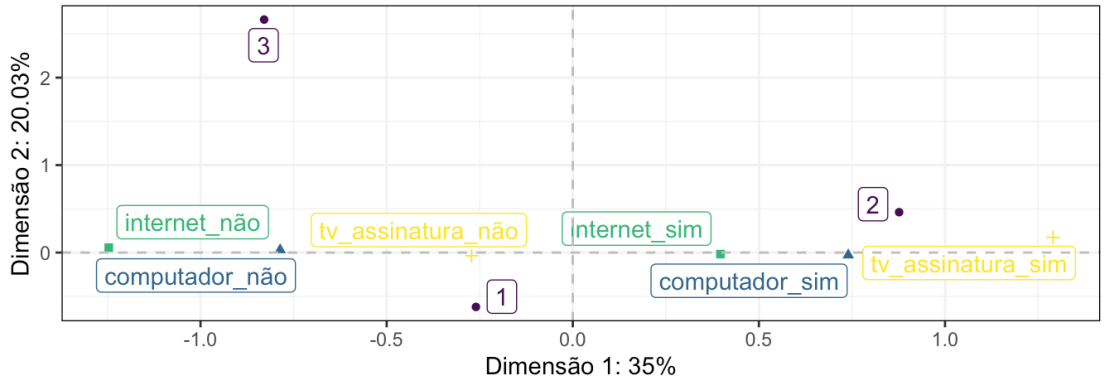
Relatórios gerados – Análise de Agrupamentos e ACM

Estado de Roraima
Materiais e métodos
Relatórios (R markdown) disponíveis em https://bit.ly/AW_TCC_ENEM-2020



Agrupamentos
Variáveis sócio-econômicas:
- TV por assinatura
- Computador em casa
- Internet em casa

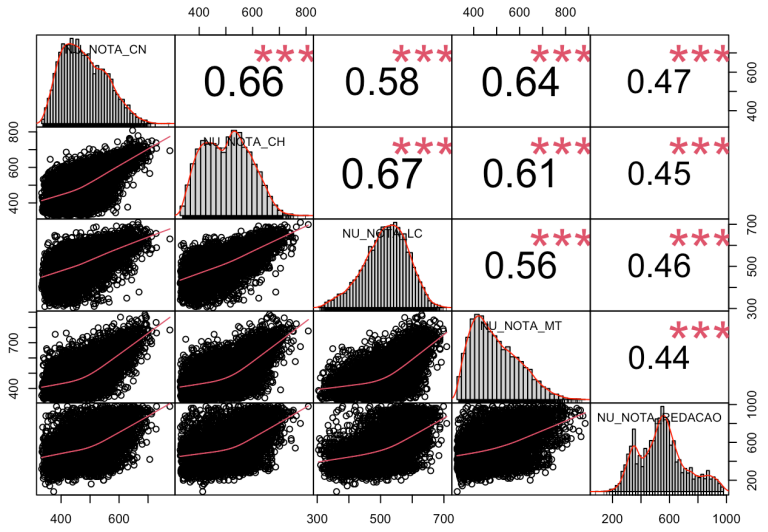


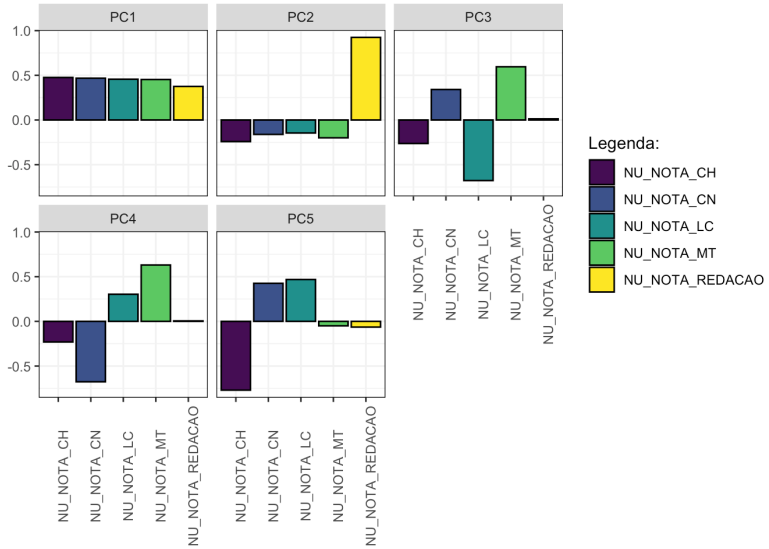


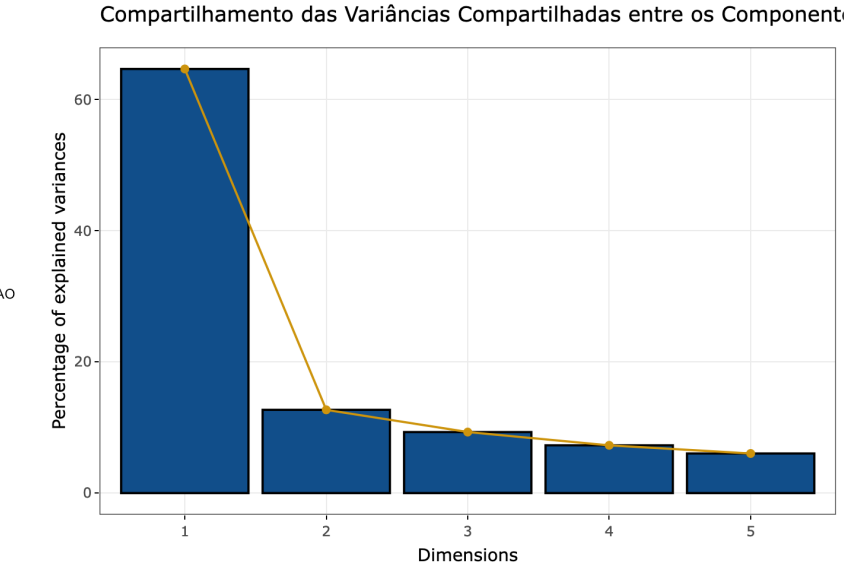
O objetivo desta análise foi avaliar a relação do desempenho no ENEM e algumas variáveis sócio-econômicas

Relatórios gerados – PCA

Estado de Roraima
Materiais e métodos
Relatórios (R markdown) disponíveis em https://bit.ly/AW_TCC_ENEM-2020



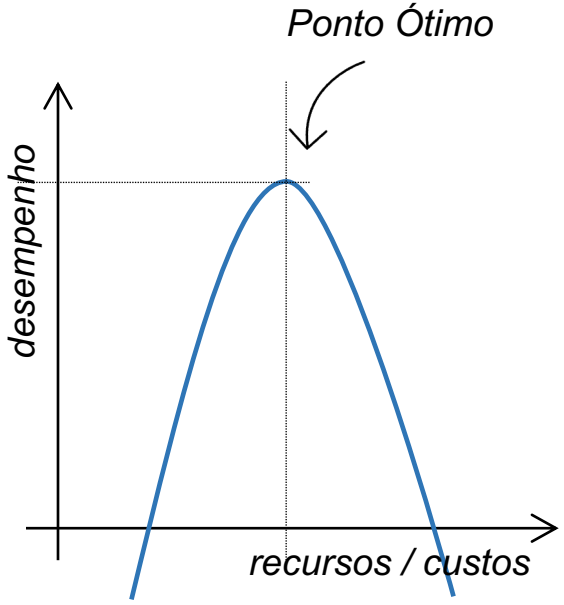




	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT	NU_NOTA_REDACAO	Fator1	pontuacao
	540.6	584.0	609.4	672.5	900	1.6497888	1.0669655
	586.5	650.9	568.3	639.0	800	1.6479832	1.0657978
	604.3	585.5	548.2	626.9	500	1.0360876	0.6700674
	498.3	518.9	485.6	396.8	560	-0.2224172	-0.1438435
	461.5	454.1	455.7	382.7	640	-0.5811558	-0.3758501

O objetivo desta análise foi criar um ranking de desempenho dos participantes, para verificação de potencial impacto de cada nota no desempenho geral.

Os resultados…



- ✓ Cada algoritmo (e etapa de processamento) tem uma necessidade diferente de recursos.
- ✓ Diferentes etapas de processamento = necessidades diferentes de recursos. **Dimensionamento deve considerer cada e todas as fases**.
- ✓ **Buscar o "Ponto Ótimo"**.
  - ✓ Alocar mais recursos (memória ou CPU) que o necessário não garante melhor tempo de processamento, mas aumenta custo.
  - ✓ Alocar menos memória pode induzir ao uso de mais disco físico e CPU (swapping), e se não houver recursos adicionais para isso, erros e falhas no processamento.
- ✓ Há **outros fatores** (além de configuração de memória, CPU e disco) que afetam o desempenho do algoritmo.
- ✓ Ambientes com **configuração semelhantes tendem a gerar resultados semelhantes** (cloud vs on-premise).
- ✓ Ambientes em cloud permitem **escalabilidade** dos recursos, sem necessidade de novos grandes investimentos, ou tempo de espera.

Os resultados…

Tipo de ambiente (PC x CSP) trazem escalas diferentes de custo total – CAPEX x OPEX

Uso híbrido pode ser opção → início on-premise, e à medida mais capacidade é necessária, usa-se escalabilidade da cloud

CAPEX maior ⟶ **PC / On-Premise** – Custo apenas do equipamento.
Outros custos a considerar: eletricidade, manutenção, outros equipamentos, upgrade/reposição do equipamento

OPEX maior ⟶ **CSP (Cloud Service Provider)** – Custo apenas do uso.
Outros custos a considerar: acesso à internet (provedor de acesso e equipamento pessoal)

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

Conversão US$ → R$, à taxa de câmbio comercial em 30 set. 2022 (R$ / US$: 5,406)

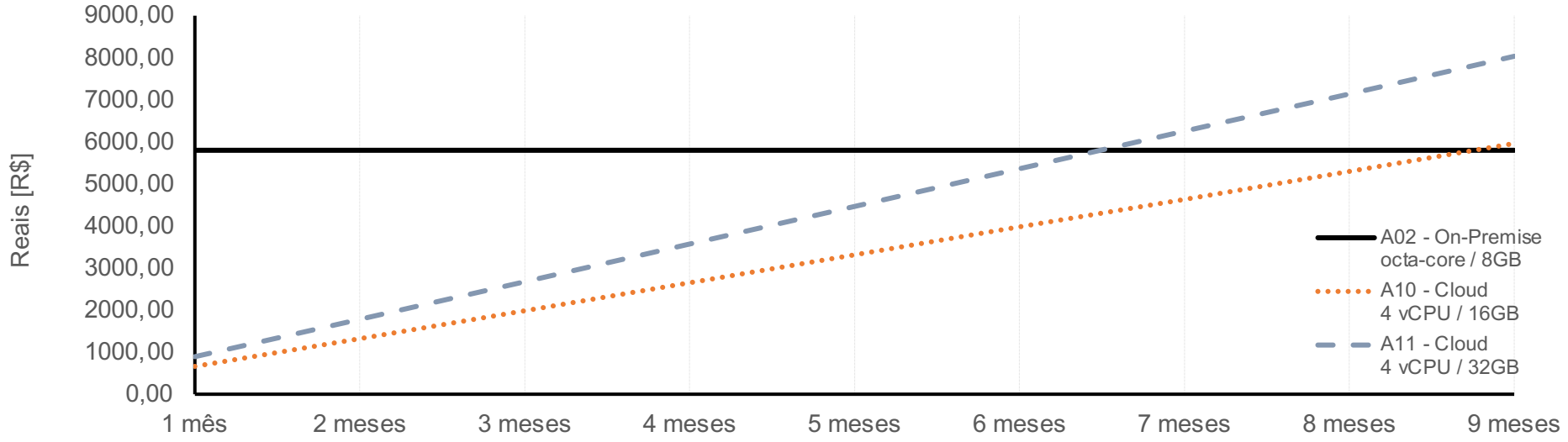
Os resultados…

O custo total on-premise (PC) é único e assim, constante ao longo do tempo.*

O custo total na cloud aumenta conforme uso, e após algum tempo, esse custo acumulado ultrapassa on-premise

*ATENÇÃO!

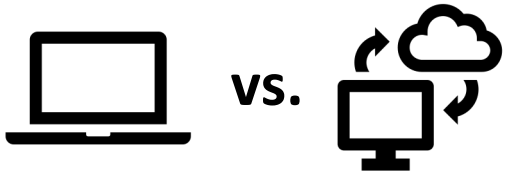
ambiente	custo / h	1 mês	2 meses	3 meses	4 meses	5 meses	6 meses	7 meses	8 meses	9 meses
A02	5800,00	5800,00	5800,00	5800,00	5800,00	5800,00	5800,00	5800,00	5800,00	5800,00
A10	0,92	662,40	1324,80	1987,20	2649,60	3312,00	3974,40	4636,80	5299,20	5961,60
A11	1,24	892,80	1785,60	2678,40	3571,20	4464,00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



Projeção de Custos acumulados dos ambientes com melhor performance
Assume-se ambiente em nuvem ativo sem interrupção / parada manual dos recursos

Considerações finais…

… em função das escolhas



Necessidade e avaliações individuais:

- ✓ Algoritmos em uso e técnicas de programação
- ✓ Capacidade de investimento inicial / longo prazo
- ✓ “Cloud”: CAPEX mínimo, cuidados (gastos, dimensionamento correto).
- ✓ “On-Premise”: CAPEX alto, responsabilidade por infraestrutura
- ✓ Uso híbrido – simultâneo ou complementar

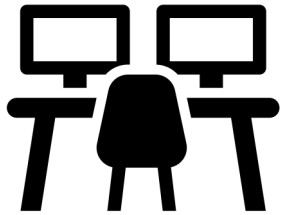
Outras avaliações:

- ➢ Outros parâmetro (algorítimos, dados, objetivo de análise)
- ➢ Outros critérios (segurança, latência, linguagem)
- ➢ Outros ambientes (provedores, cenários/ferramentas de ML)

Extensões do Estudo ou Estudos futuros

Considerações finais…

… aplicações práticas das escolhas



Profissional / Pequena Empresa

Usos:
- Cloud
- Híbrido

- ✓ Poucos recursos para investimento inicial.
- ✓ Uso de recursos já adquiridos, ou com menor capacidade para início e crescimento posterior.

Educação – Professores / Estudantes

Usos:
- Cloud

- ✓ Padronização dos ambientes – professor / alunos
- ✓ Padronização dos acessos (facilita distribuição de conteúdo/software e determinação de erros)

Profissionais / Indivíduos

Usos:
- "On-premise"
- Cloud
- Híbrido

- ✓ Acesso aos dados e análise com equipamento pessoal; e mesmo na ausência de equipamento robusto, acesso pela cloud ou híbrido

Obrigada!

Ao meu marido Ricardo

Ao meu filho Marcos Vinícius

À minha mãe Dalney

Aos meus amigos e colegas

Ao professor Thiago

Aos professores e monitores do curso

À equipe Pecege

Aos professores da banca

A todos assistindo essa defesa.

Adriana Weingart

about.me/adrianaweingart

Slides Auxiliares

Provedores de Serviço de “Cloud”

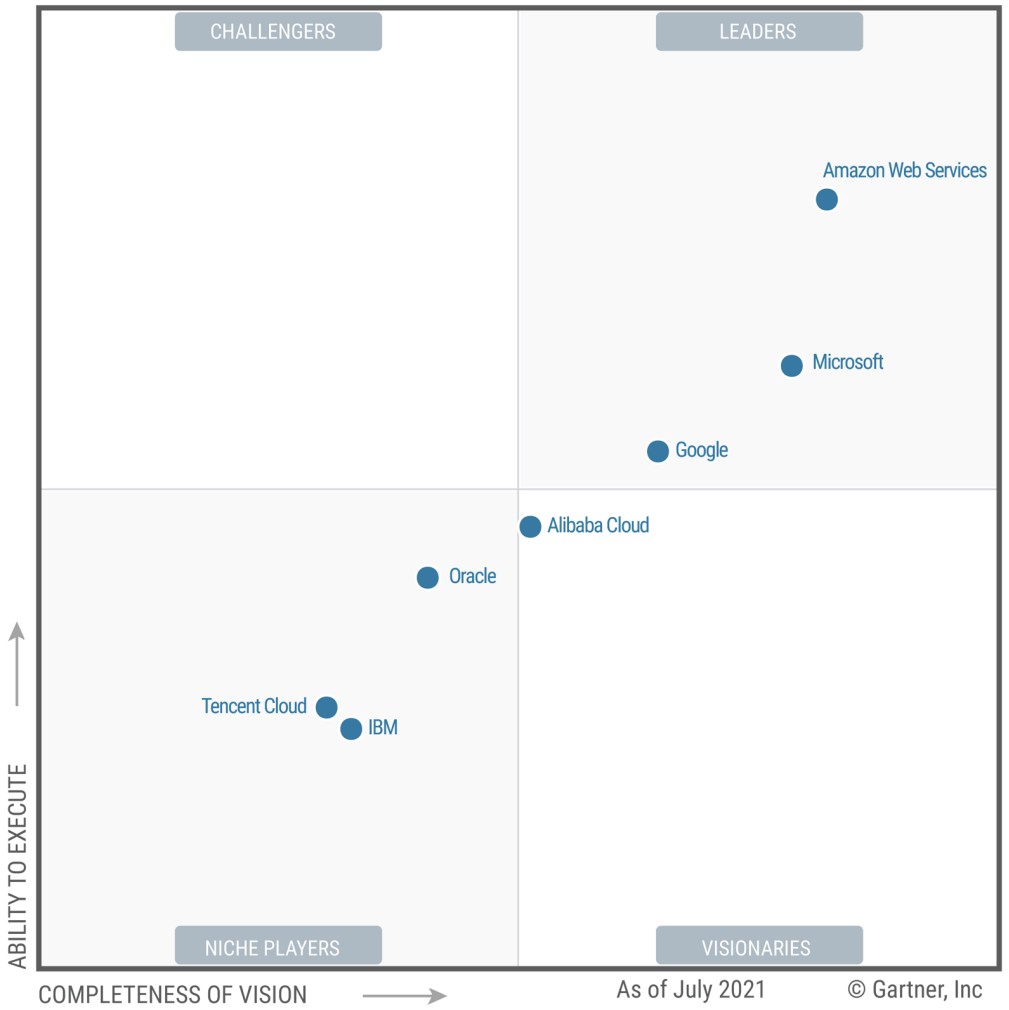
Materiais e métodos

“Magic Quadrant for Cloud Infrastructure and Platform Services”

Amazon Web Services [AWS]

Microsoft Azure

Google Cloud Platform [GCP]



Source: Gartner (July 2021)

Fonte: https://www.gartner.com/doc/reprints?id=1-271OE4VR&ct=210802&st=sb

Critérios para análise e comparação

Materiais e métodos

- ✓ Custo
- ✓ Segurança
- ✓ Performance

31 atributos de qualidade de serviço

Criteria for Selecting Cloud Service Providers: A Delphi Study of Quality-of-Service Attributes
Lang, M.; Wiesche, M.; Krcmar, H. 2018

Critérios mais aborados em 77 artigos publicados

Cloud service evaluation method-based Multi-Criteria Decision-Making: A systematic literature review
Alabool, H.; Kamil, A.; Arshad, N.; Alarabiat, D. 2018

Ambientes de testes

Materiais e métodos

(*) Equipamento pessoal

#	equipamento	provedor	SO e versão	Proc	Mem	Disco	R	RStudio	Região, Custo / h
A01	MacBook Pro	n/a	maOS Monterey	Dual Core	8 GB	256 GB	4.2.1	2022.07.1 build 554	R$15.300,00 (*)
A02	Dell Vostro 361	n/a	Windows 10 Pro	i7-10700 (8-core)	8 GB	512 GB	4.2.1	2022.07.2 build 576	R$5.800,00 (*)
A03	t2.micro	AWS	Ubuntu 18.04	1 vCPU	1 GB	30 GB	4.0.2	Server 1.3.1073	us-central1, free
A04	Plano cloud free	RStudio Cloud	Ubuntu 20.04	1 vCPU	1 GB	n/a	4.2.1	2022.02.2 build 485	free
A05	e2-standard-2	GCP	Ubuntu 18.04	2 vCPU	8 GB	10 GB	4.2.1	2022.07.1 Build 554	us-central1, US$0,07
A06	e2-standard-2	GCP	Ubuntu 18.04	2 vCPU	8 GB	70 GB	4.2.1	2022.07.1 Build 554	us-central1, US$0,08
A07	e2-standard-8	GCP	Ubuntu 18.04	8 vCPU	32 GB	70 GB	4.2.1	2022.07.1 Build 554	us-central1, US$0,28
A08	t2.large	AWS	Ubuntu 18.04	2 vCPU	8 GB	30 GB	4.0.2	Server 1.3.1073	us-east1, US$0,0928
A09	Standard_D2ads_v5	Azure	Ubuntu 18.04	2 vCPU	8 GB	8 GB	3.4.4	Server 1.3.1093	West US 3, US$0,086
A10	Standard_D4as_v5	Azure	Ubuntu 18.04	4 vCPU	16 GB	16 GB	3.4.4	Server 1.3.1093	West US 3, US$0,17
A11	Standard_E4as_v5	Azure	Ubuntu 18.04	4 vCPU	32 GB	32 GB	3.4.4	Server 1.3.1093	West US 3, US$0,23
A12	t2.2xlarge	AWS	Ubuntu 18.04	8 vCPU	32 GB	30 GB	4.0.2	Server 1.3.1073	us-east1, US$0,23

Dados e algoritmos para teste

Materiais e métodos
Scripts disponíveis em https://bit.ly/AW_TCC_Scripts

(1) Amostra normal de 5.000.000 entradas
(2) Microdados do ENEM 2020 – arquivo .csv de 2,02GB com 76 variáveis e 5.783.109 observações

#	Amostra	Modelo	Pacotes usados
T1	Amostra normal gerada (1)	Modelo de regressão simples	stats, tictoc
T2	Microdados do ENEM 2020(2)	Análise Exploratória dos dados	tidyverse, stringi, gridExtra, kableExtra, psych, tictoc, knitr
T3	Microdados do ENEM 2020(2)	Análise de Agrupamentos e Análise de Correspondência Múltipla [ACM]	tidyverse, kableExtra, knitr, tictoc, factoextra, cabootcrs, FactoMineR, ggrepel, gridExtra
T3r	Microdados do ENEM 2020 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(2)	Análise Fatorial por Componentes Principais [PCA]	tidyverse, kableExtra, knitr, tictoc, reshape2, psych, PerformanceAnalytics, ggrepel, plotly, factoextra
T4r	Microdados do ENEM 2020 do Estado de Roraima	Análise Fatorial por Componentes Principais [PCA]	tidyverse, kableExtra, knitr, tictoc, reshape2, psych, PerformanceAnalytics, ggrepel, plotly, factoextra

Dados e algoritmos para teste

Materiais e métodos
Scripts disponíveis em https://bit.ly/AW_TCC_Scripts

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"
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

Ferramentas de visualização

Materiais e métodos

MS Excel

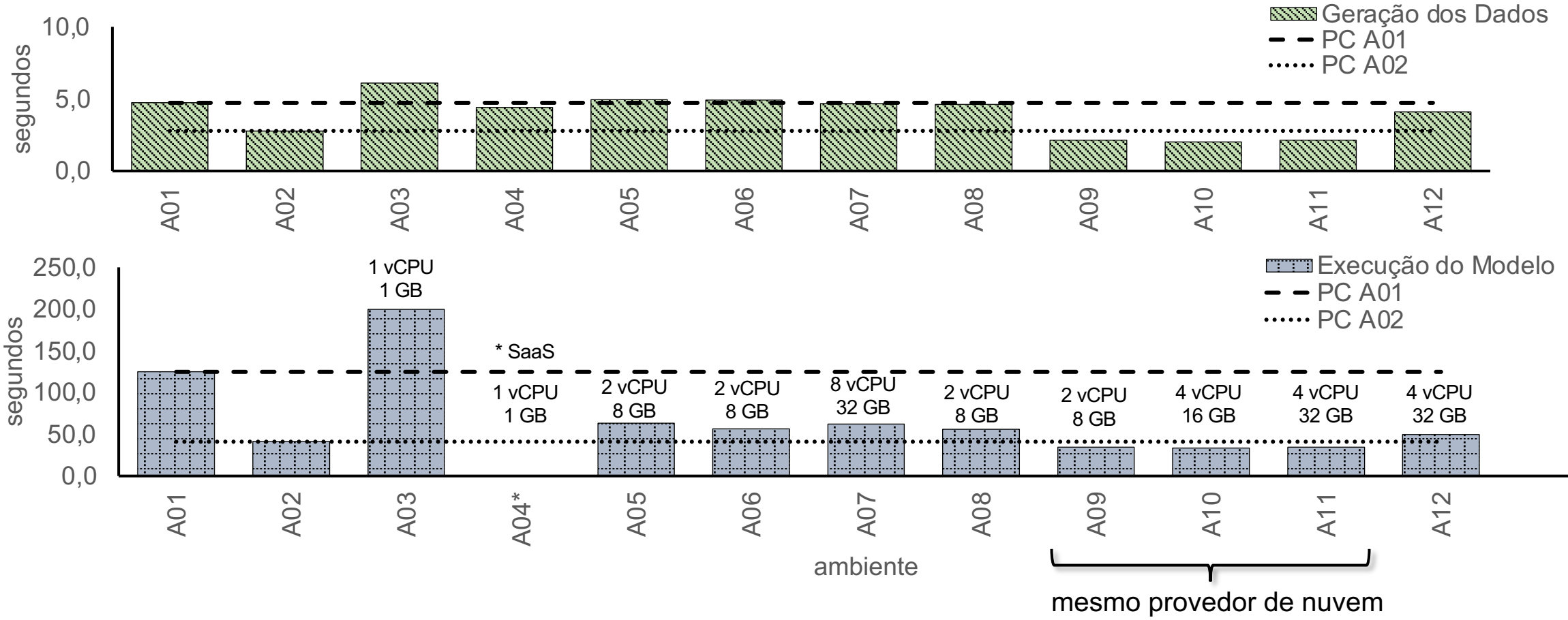
Coblis —
Color Blindness Simulator
(ferramenta online)

Os resultados…

Alocar mais "memória" que o necessário não garante melhora no processamento

Outros fatores também influem no desempenho da execução do algoritmo

Teste T1



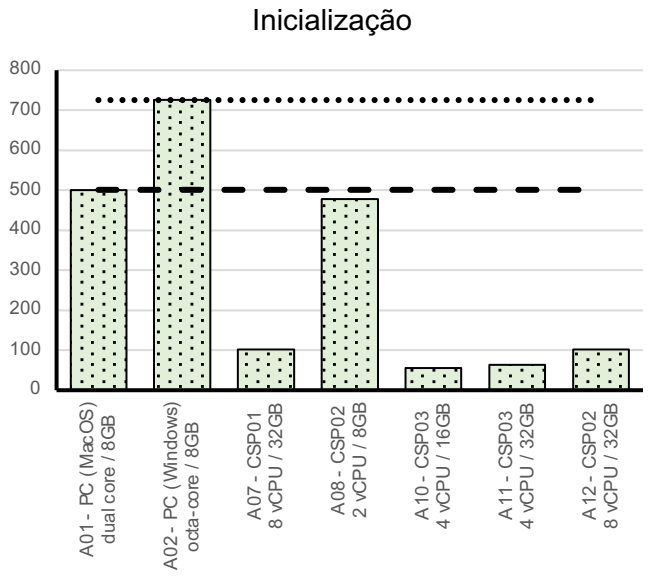
Os resultados…

Alocação insuficiente de recursos (memória e/ou memória + disco) pode não permitir análise – não importa ambiente

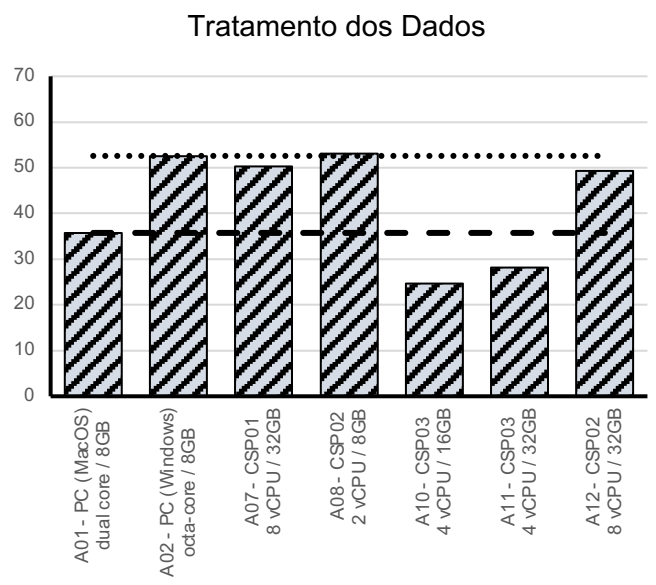
Alocar mais "memória" que o necessário não garante melhora no processamento

Etapas diferente do processamento = necessidades diferentes. Dimensionamento deve considerar cada e todas as fases

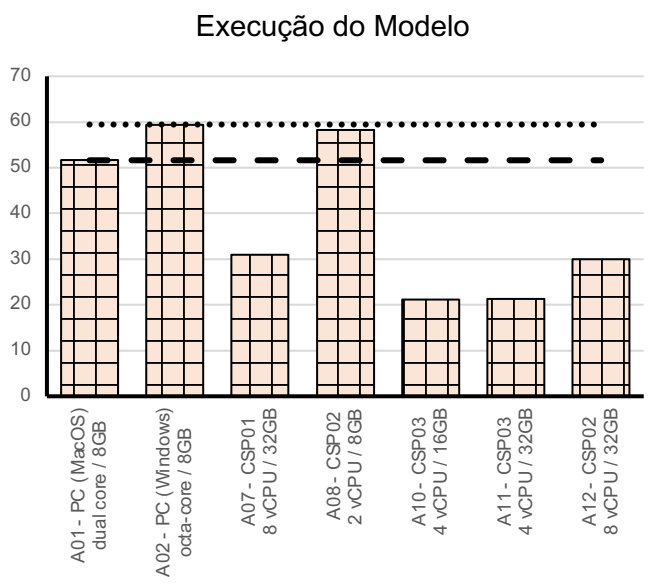
Teste T2



Carregamento dos dados
Mais recursos (CPU e Memória) reduziram significantemente o tempo de processamento



Tratamento dos dados (clean up e redução do dataset)
Provedor de Cloud específico, com melhor capacidade nas VMs tiveram melhor tempo

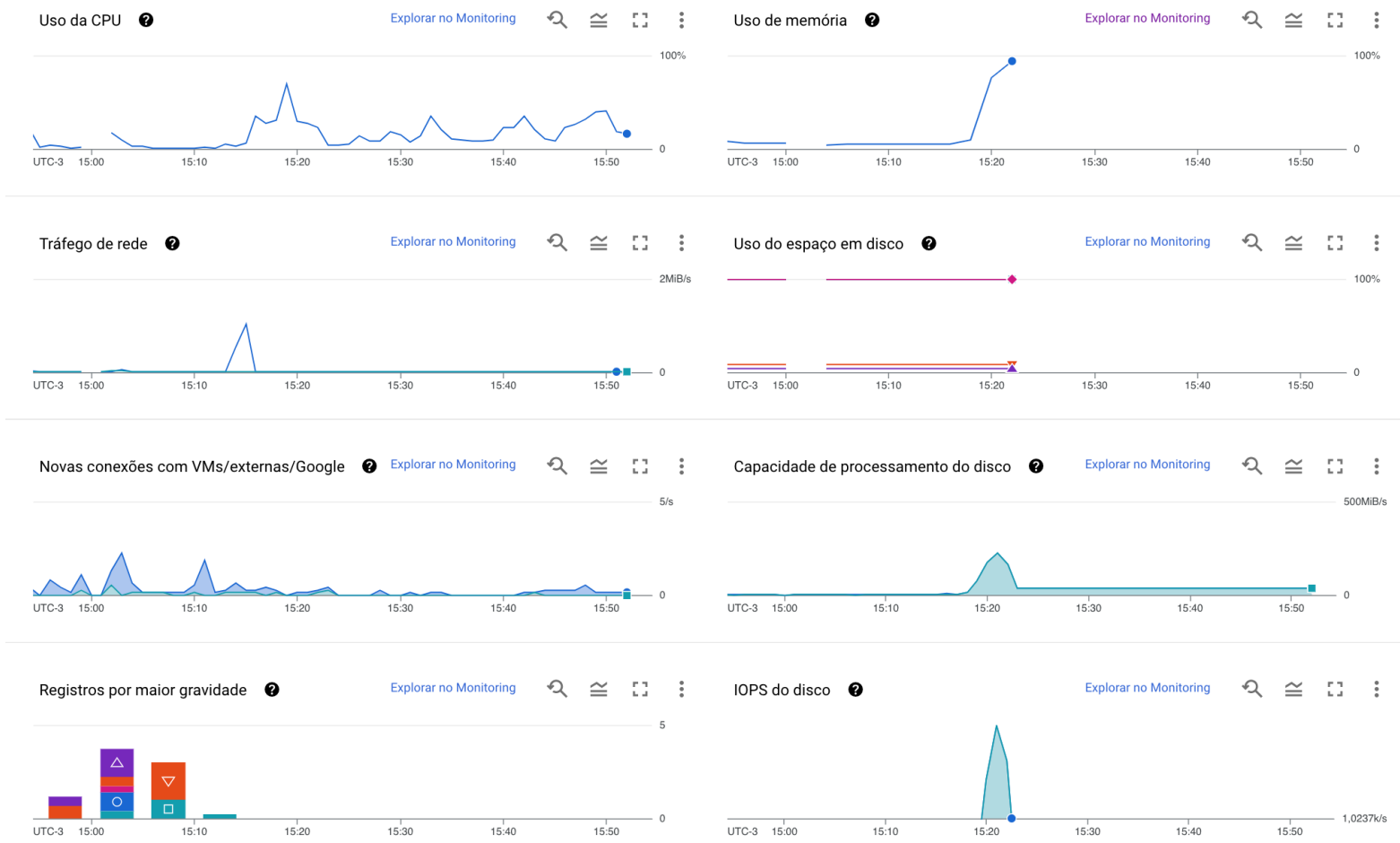


Elaboração de gráficos / visualização
Recursos disponíveis em cloud apresentaram melhor resultados

**A05 e A06 falharam após SWAP*

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

Teste T2



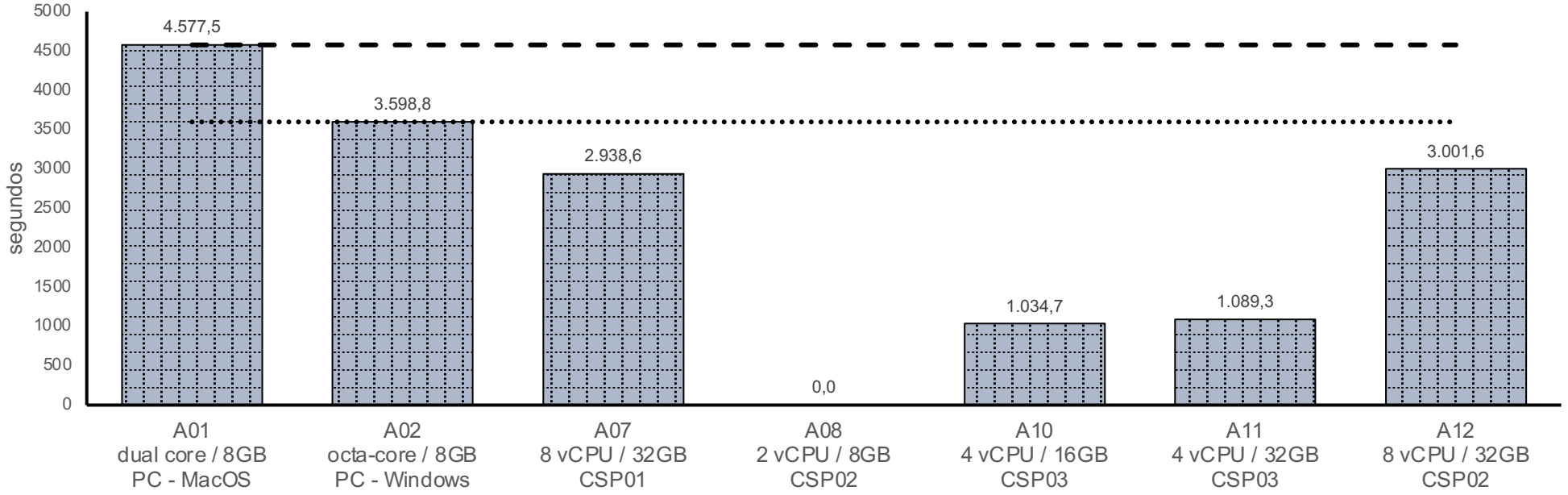
Os resultados…

Entender as necessidades do algoritmo para correta alocação dos recursos

→ mais CPU não significa melhor performance se o algoritmo requer mais memória.

Ambientes em cloud permitem escalabilidade dos recursos, sem necessidade de novos grandes investimentos

Teste T3



Os resultados…

Ambientes com configuração semelhante geram resultados semelhantes (cloud vs on-premise)

Adicionar mais recursos (CPU e memória) não garantem melhoria no resultado – sizing deve ser correto

Teste T4