

# ANÁLISE DE DADOS UTILIZANDO CLUSTER DE BAIXO CUSTO

## COMPARAÇÃO DE DESEMPENHO DE AMBIENTES VIRTUAIS

Felipe Fonseca Rocha

Orientador: Ítalo Fernando Scotá Cunha

Universidade Federal de Minas Gerais

09 de Fevereiro de 2022



## Sumário

### 1 Objetivo

- ▶ Objetivos específicos

### 2 Introdução

- ▶ Motivação
- ▶ Justificativa
- ▶ Abordagem

### 3 Revisão de literatura

- ▶ Análise de dados
- ▶ Alternativas open source
- ▶ Cluster orquestrador de container

### 4 Método

- ▶ Disponibilidade dos recursos deste trabalho
- ▶ Especificação dos nós integrantes cluster de baixo custo
- ▶ Plataforma de orquestração de carga de trabalho
- ▶ Configuração e provisionamento do cluster
- ▶ Análise de dados
- ▶ Monitoramento

# 1 Objetivo

- ▶ Objetivos específicos

## 2 Introdução

## 3 Revisão de literatura

## 4 Método

## 5 Conclusão

## Objetivo

Realizar a comparação de desempenho de orquestração de recursos em cluster de baixo custo em ambientes virtualizados: completa, Sistema Operacional; para o processamento e a análise de dados em saúde, como tendência de consumo de azitromicina entre os anos 2014 e 2021.

- Realizar a orquestração de recursos em cluster de baixo custo;
- Comparar o desempenho de clusters em ambientes virtualizados;
- Validar o uso de um cluster de utilização compartilhada para processamento de dados distribuídos;
- Propor um método de análise em cluster Kubernetes® com uso de computadores desktops;

## 1 Objetivo

## 2 Introdução

- ▶ Motivação
- ▶ Justificativa
- ▶ Abordagem

## 3 Revisão de literatura

## 4 Método

## 5 Conclusão

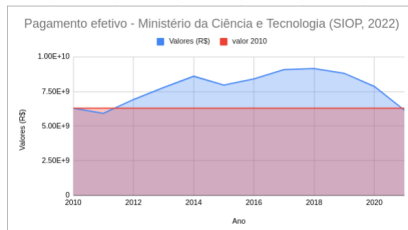


## Introdução – Motivação

- Uso de ferramentas de análise dados em saúde (GALVÃO; VALENTIM, 2019)
- Integração de Sistemas de Informação em Saúde e necessidade de facilitar processo de análise de grandes volumes de dados (GALVÃO; VALENTIM, 2019; MEHTA; PANDIT, 2018a)
- Disponibilidade de dados pelo Decreto nº 8.777 (Brasil, 2016) e a necessidade de extração de informações.

## Introdução – Justificativa

- Restrição Orçamentária
  - Diminuição de verbas para ciência e tecnologia –2,32%
  - Aumento do dólar em mais de 3,27%
- Disponibilidade de estratégias de análise





## Introdução – Justificativa

- Restrição Orçamentária
  - Diminuição de verbas para ciência e tecnologia –2,32%
  - Aumento do dólar em mais de 3,27%
- Disponibilidade de estratégias de análise
- Tomada de decisão em saúde, mais de 152mi de Brasileiros dependem exclusivamente do SUS (UNASUS, 2021)



## Introdução – Abordagem

- Cluster Kubernetes®
- Cargas de trabalho – Analise de tendencia de uso de azitromicina entre 2014 e 2021
- 2 abordagens de virtualização:
  - completa – *Hypervisor* tipo 2
  - sistema operacional – contêineres
- Máquinas comuns e de baixo poder computacional:
  - 1 vCPU
  - 2 GB de RAM
  - 6-8 máquinas

## Introdução – Abordagem

- Aplicação de abordagem DevOps:
  - *Shift Right* – Fazes finais do *SDLC*
  - CI (integração contínua) e CD (entrega contínua) – deploy da aplicação e início do monitoramento
  - IaC (infraestrutura como código) – acuracia na repetição dos procedimentos de provisionamento de recursos e configuração.
- Uso de metodos do tipo USE (utilização, saturação e erro) para comparação entre as cargas de trabalhos e ambientes de simulação e virtualização

## 1 Objetivo

## 2 Introdução

## 3 Revisão de literatura

- ▶ Análise de dados
- ▶ Alternativas open source
- ▶ Cluster orquestrador de container

## 4 Método

## 5 Conclusão



## Revisão de literatura- Análise de dados

- Complexidade de tomar decisão em saúde (ANDRADE, 2008; RESENDE; VIANA; VIDGAL, 2009)
- Definição de Big Data 5 Vs, complexidade e Destruturação (LANEY et al., 2001; DIJCKS, 2013; CORPORATION, 2012)
- Complexidade de relacionar dados por multifatoriedade (FACELI et al., 2011)
- Não consolidação de métodos de uso de Big Data em saúde, especialmente em estudos quantitativos, foco em custo e decisões clínicas (MEHTA; PANDIT, 2018b)

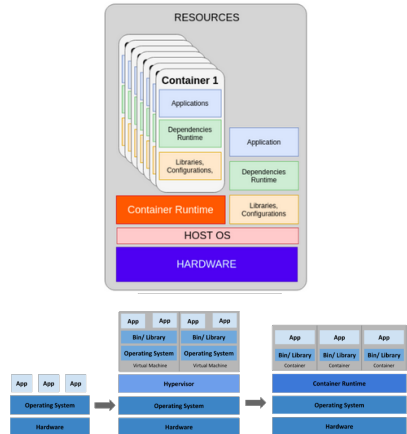
## Revisão de literatura- Alternativas open source

- Agrupamento em categorias:
  - Computação em nuvem privada:
    - ▶ Tecnologias: OpenStack®, CloudStack®
    - ▶ Requisitos exigentes
    - ▶ Complexidade: SaaS, PaaS e SaaS (OPENSTACK, 2022; CLOUDSTACK, 2022; MELL; GRANCE, 2011)

## Revisão de literatura- Alternativas open source

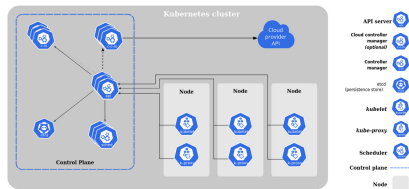
- Orquestração de Containers:

- Kubernetes®
- Apache Mesos®
- Hashicorp Nomad®
- Docker Swarm®



# Revisão de literatura- Cluster orquestrador de container

- Kubernetes®:
  - Origem de 15 anos de trabalho da Google (Borg) (VERMA et al., 2015)
  - estrutura de objetos compostos (KUBERNETES, )
    - ▶ Kube-apiserver
    - ▶ Kube-scheduler
    - ▶ Kube-controller-manager
    - ▶ Kubelet
    - ▶ Kube-proxy
    - ▶ Pod





## 1 Objetivo

## 2 Introdução

## 3 Revisão de literatura

## 4 Método

- ▶ Disponibilidade dos recursos deste trabalho
- ▶ Especificação dos nós integrantes cluster de baixo custo
- ▶ Plataforma de orquestração de carga de trabalho
- ▶ Configuração e provisionamento do cluster
- ▶ Análise de dados
- ▶ Monitoramento
- ▶ Comparação entre tipos de virtualização
- ▶ Cronograma

## 5 Conclusão

---

---

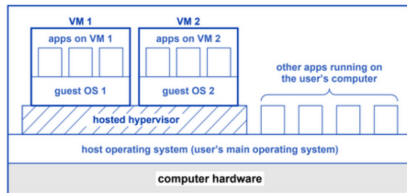
---

## Método – Disponibilidade dos recursos deste trabalho

Todos os componentes definidos nesse trabalho estarão contidos em um ou mais repositórios públicos, garantindo assim a livre apreciação da comunidade não só científica, mas a todos os interessados na contribuição ou utilização sob a licença pública geral GNU versão 3 (GNU, ). <https://github.com/felipefrocha/esufmg-tcc> Repositório

## Método - Especificação dos nós integrantes cluster de baixo custo

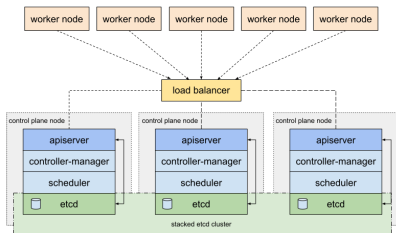
- Cluster Simulado:
  - Virtualização:
    - ▶ Maquinas Virtuais (VMs) (*Hypervisor* tipo 2)
    - ▶ Contêineres Aninhados (Docker In Docker, ou DinD)
  - Especificações de hardware  
1vCPU, 2 GB de RAM;
- provisionamento em 2 etapas
- máquinas subutilizadas
- CAPEX



# Método - Plataforma de orquestração de carga de trabalho

- Multi-master Etcd atachado (KUBERNETES, ; ETCD, )
- Arquitetura sugerida para produção
- Alta disponibilidade do cluster
- Recursos limitados

kubeadm HA topology - stacked etcd



## Método – Plataforma de orquestração de carga de trabalho

- Implantação da carga de Trabalho
  - Container
  - Parametrizável
  - Volume compartilhado
- Ciclo de vida da aplicação
  - Monorepo (BRITO; TERRA; VALENTE, 2018)
  - Padronização de código
  - Sincronização
  - Deploy em Pipeline

## Método – Configuração e provisionamento do cluster

- Gerenciador de Configuração (Ansible®)
- *Agentless*
- Idempotência
- Gerenciamento de inventário
- SSH – Escolha do algoritmo de criptografia (RFC4254, )

## Método – Análise de dados

- “Vendas de Medicamentos Controlados e Antimicrobianos – Medicamentos Industrializados”
- $530 \cdot 10^6$  linhas com mais de 70 GB
- Análise de tendência do consumo de azitromicina
- Caso base – comparação com processo de análise em *bare metal* 8vCPU, 16 GB de RAM – totalizando o poder computacional total do cluster proposto

## Método – Monitoramento

- *OpenTelemetry*
- *Prometheus* Monitoramento de sistemas e Banco de dados de series temporais
- *Grafana* – Dashboard e observabilidade
- Parametros de tempo, taxa de utilização de memoria e processamento





## Método – Comparação entre tipos de virtualização

- A arquitetura x86 comum e maior poder computacional (FAYYAD; LucPerneel; TIMMERMAN, 2013)
- macrobenchmark (system level benchmark) (HUGE, ; SCHEEPERS, 2014)
- Parametros de tempo, taxa de utilização de memoria e processamento
- Maquina hospedeira e virtuais serão avaliadas durante o processamento
- Metodo USE de avaliação (GREGG, )
- APM Application Performance Management) associada a aplicação da carga de trabalho por *OpenTelemetry* (TANG et al., 2021)

		Ondas			Dútes		Cascata	
Fases de Projeto	Atividades do TCC	Atividades	Objetivos	S	Data Inicial	Data Final		
Exploratório	Proposta TCC I	Elaboração de estratégias de busca	Identificar estudos parecidos, explorar tecnologias disponíveis e avaliar oportunidades e conceitos associados aos usuários	1	17/10/2021	23/10/2021		
		Busca e avaliação dos artigos selecionados		2	24/10/2021	30/10/2021		
		Escrita de resumo bibliográfico		3	31/10/2021	06/11/2021		
Concepção	Visão Geral do Projeto	Descrição formal dos stakeholders	Identificar público alvo, validar ideia da solução e listar alternativas	7	07/11/2021	04/12/2021		
		Avaliação de alternativas		9	05/12/2021	18/12/2021		
		Elaboração da fundamentação teórica e justificativa		10	19/12/2021	25/12/2021		
Desenvolvimento	Marcação de Defesas Teoria Inicial Monografia & Versão Final da Monografia	Especificação e critérios de aceitação	Elaborar detalhamento da solução, mapear fronteiras da solução, identificar riscos ao projeto e propor desenho inicial da solução	12	26/12/2021	08/01/2022		
		Levantamento de Requisitos		13	09/01/2022	15/01/2022		
		Levantamento de Lista de Materiais e softwares		15	16/01/2022	29/01/2022		
		Apresentação do estudos e resultados de PoCs		17	30/01/2022	12/02/2022		
		Avaliação de viabilidade do sistema		20	13/02/2022	05/03/2022		
Produção	TCC II	Implementação da montagem (caso viável) e testes de verificação	Produção, Inspeção, Verificação e Validação da solução proposta.	21	06/03/2022	12/03/2022		
		Instrumentação (software) e verificação		26	13/03/2022	16/04/2022		
		Implementação da análise e verificação		28	17/04/2022	30/04/2022		
Utilização & Suporte		Testes de Validação	Captação da utilização em cenário real em projeto de pesquisa parceiro	30	01/05/2022	14/05/2022		
		Coleta dos resultados		31	15/05/2022	21/05/2022		
Encerramento		Discussão dos resultados obtidos	Estudo do caso de uso e sumarização dos resultados para apresentação da solução junto a banca	33	22/05/2022	04/06/2022		
		Definição de próximas etapas		34	05/06/2022	11/06/2022		
		Formalização dos trabalhos e apresentação		35	12/06/2022	18/06/2022		

1 Objetivo

2 Introdução

3 Revisão de literatura

4 Método


5 Conclusão


## Conclusão

- Avaliação de diferentes tipos de virtualização
- Seleção de plataforma de orquestração de cargas de trabalhos com base em requisitos e restrições
- Análise dos impactos socio-econômicos oriundos da restrição orçamentária a pesquisa de uma forma geral
- Desenho de uma estratégia de extração de informações relevantes de uma base de dados com volume considerável
- Entendimento da complexidade dos fatores considerados no processo de decisão em saúde


Trabalhos futuros contemplarão a implementação, testes e coletas de dados para avaliação comparativa das virtualizações propostas no ambiente simulado. Baseado nesses resultados pode se evoluir essa discussão na forma de recrutamento de computadores para o cluster de maneira a garantir o isolamento da máquina base.


## Referências I

 ANDRADE, A. Q. d. A tomada de decisão e sistemas de informação em saúde. jan. 2008. Accepted: 2019-08-13T14:49:40Z Publisher: Universidade Federal de Minas Gerais. Disponível em: [⟨https://repositorio.ufmg.br/handle/1843/ECIC-7XMFGC⟩](https://repositorio.ufmg.br/handle/1843/ECIC-7XMFGC).

 Brasil. **DECRETO Nº 8.777, DE 11 DE MAIO DE 2016**. 2016.  
[⟨http://www.planalto.gov.br/ccivil\\_03/\\_ato2015-2018/2016/decreto/d8777.htm⟩](http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm).

 BRITO, G.; TERRA, R.; VALENTE, M. T. Monorepos: A Multivocal Literature Review. **arXiv:1810.09477 [cs]**, out. 2018. ArXiv: 1810.09477. Disponível em: [⟨http://arxiv.org/abs/1810.09477⟩](http://arxiv.org/abs/1810.09477).

 CLOUDSTACK. **Installation Guide — Apache CloudStack 4.16.0.0 documentation**. 2022. Disponível em: [⟨https://docs.cloudstack.apache.org/en/latest/installguide/index.html⟩](https://docs.cloudstack.apache.org/en/latest/installguide/index.html).

 CORPORATION, I. **Big Data Analytics: Intel's IT Manager Survey on How Organizations Are Using Big Data**. [S.l.], 2012.

---

---

---

## Referências II



DIJCKS, J.-P. **Oracle: Big data for the enterprise**. [S.l.], 2013.



ETCD. **Install**. Section: docs. Disponível em: <<https://etcd.io/docs/v3.5/install/>>.



FACELI, K. et al. Inteligência artificial: uma abordagem de aprendizado de máquina. 2011.




FAYYAD, H.; LucPerneel; TIMMERMAN, M. Benchmarking the Performance of Microsoft Hyper-V server, VMware ESXi and Xen Hypervisors. **Journal of Emerging Trends in Computing and Information Sciences**, Vol. 4, No. 12 , December 2013, pp: 922-933, ISSN 2079-8407, dez. 2013.





GALVÃO, A. B.; VALENTIM, R. A. d. M. Desafios para os Avanços da Análise de Big Data na Saúde. In: **Anais Estendidos do Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)**. SBC, 2019. p. 155–160. ISSN: 2763-8987. Disponível em: <[https://sol.sbc.org.br/index.php/sbcas\\_estendido/article/view/6301](https://sol.sbc.org.br/index.php/sbcas_estendido/article/view/6301)>.


## Referências III

 GNU. **A Quick Guide to GPLv3 - GNU Project - Free Software Foundation.** Disponível em: [⟨https://www.gnu.org/licenses/quick-guide-gplv3.en.html⟩](https://www.gnu.org/licenses/quick-guide-gplv3.en.html).

 GREGG, B. **The USE Method.** Disponível em: [⟨https://www.brendangregg.com/usemethod.html⟩](https://www.brendangregg.com/usemethod.html).


 HUGE, M. **Different Types of Benchmarks.** Disponível em: [⟨https://www.cs.umd.edu/users/meesh/cmsc411/website/projects/morebenchmarks/types.html⟩](https://www.cs.umd.edu/users/meesh/cmsc411/website/projects/morebenchmarks/types.html).

 KUBERNETES. **Kubernetes Documentation | Kubernetes.** Disponível em: [⟨https://kubernetes.io/docs/home/⟩](https://kubernetes.io/docs/home/).


 LANEY, D. et al. 3d data management: Controlling data volume, velocity and variety. **META group research note**, Stanford, v. 6, n. 70, p. 1, 2001.


## Referências IV

 MEHTA, N.; PANDIT, A. Concurrence of big data analytics and healthcare: A systematic review. **International Journal of Medical Informatics**, v. 114, p. 57–65, jun. 2018. ISSN 1386–5056. Disponível em: [〈https://www.sciencedirect.com/science/article/pii/S1386505618302466〉](https://www.sciencedirect.com/science/article/pii/S1386505618302466).

 MEHTA, N.; PANDIT, A. Concurrence of big data analytics and healthcare: A systematic review. **International Journal of Medical Informatics**, 2018.

 MELL, P.; GRANCE, T. **The NIST Definition of Cloud Computing**. [S.l.], 2011. Disponível em: [〈https://csrc.nist.gov/publications/detail/sp/800-145/final〉](https://csrc.nist.gov/publications/detail/sp/800-145/final).

 OPENSTACK. **OpenStack Docs: Xena Installation Guides**. 2022. Disponível em: [〈https://docs.openstack.org/xena/install/〉](https://docs.openstack.org/xena/install/).

 RESENDE, L.; VIANA, L.; VIDGAL, P. **PROTOCOLOS CLÍNICOS DOS EXAMES LABORATORIAIS**. 1. ed. Minas Gerais: Secretaria de Estado de Saúde de Minas Gerais, 2009.





---

---

---



## Referências V

-  RFC4254. Disponível em: [⟨https://datatracker.ietf.org/doc/html/rfc4254⟩](https://datatracker.ietf.org/doc/html/rfc4254).
-  SCHEEPERS, M. J. Virtualization and containerization of application infrastructure: A comparison. In: **21st twente student conference on IT**. [S.l.: s.n.], 2014. v. 21.
-  TANG, Y. et al. A systematical study on application performance management libraries for apps. **IEEE Transactions on Software Engineering**, IEEE, 2021.
-  VERMA, A. et al. Large-scale cluster management at Google with Borg. In: **Proceedings of the European Conference on Computer Systems (EuroSys)**. Bordeaux, France: [s.n.], 2015.

OBRIGADO  
:)

