

# ANÁLISE DE DADOS UTILIZAND *CLUSTER* E 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 Contexto e Motivação
- 2 Justificativa
- 3 Objetivo
- 4 Revisão de literatura
- 5 Método
- 6 Conclusão
- 7 Disponibilidade dos recursos deste trabalho

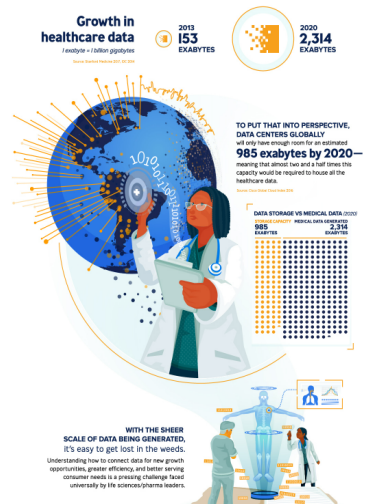
# Contexto e Motivação I

A todo momento nós geramos milhões de dados que são coletados por diferentes meios

Existem várias ferramentas disponíveis para transformá-los em informações e embasar decisões



## Contexto e Motivação II



Isso também acontece na área da saúde

Porém o uso de ferramentas de *big data* em saúde ainda é pouco significativo

Boa parte dessas ferramentas implica processamento distribuído

## Contexto e Motivação III

Potencial de melhora do sistema de saúde através de análise de dados

Integrar times com trabalho interdisciplinar

Uso de ferramentas e recursos já disponíveis de maneira correta



1 Contexto e Motivação

2 Justificativa

- Justificativa Social

3 Objetivo

4 Revisão de literatura

5 Método

6 Conclusão

7 Disponibilidade dos recursos deste trabalho

## Justificativa Social

- Tomada de decisão em saúde
- Escala: **152 milhões** dependem exclusivamente do SUS
- Restrição: Gasto de **R\$3.83** por pessoa por dia
- Volume de dados disponibilizados
- **Assertividade**
  - Ações em saúde
  - políticas públicas

1 Contexto e Motivação

2 Justificativa

- Justificativa Econômica

3 Objetivo

4 Revisão de literatura

5 Método

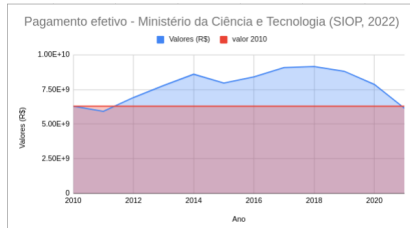
6 Conclusão

7 Disponibilidade dos recursos deste trabalho



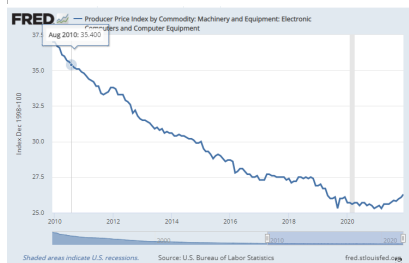
## Justificativa Econômica

- Gasto na disponibilização dos dados
- Diminuição de verbas para ciência e tecnologia -2,32%



# Justificativa Econômica

- Aumento do dólar em mais de 327% diminuindo o poder de compra
- Aumento do custo de hardware e máquinas



1 Contexto e Motivação

2 Justificativa

- Justificativa Técnica

3 Objetivo

4 Revisão de literatura

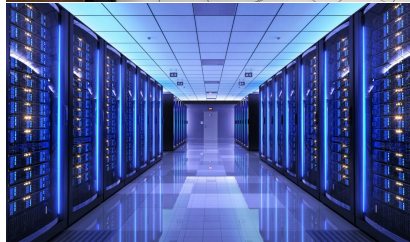
5 Método

6 Conclusão

7 Disponibilidade dos recursos deste trabalho

## Justificativa Técnica

- Necessário ser interdisciplinar
- Avaliar alternativas de processamento de dados
- Amenizar questões orçamentárias
- Melhorar uso dos recursos já existentes



# Objetivo I

## Objetivos Geral:

Realizar a comparação de desempenho de orquestração de recursos em *cluster* de baixo custo em ambientes virtualizados, para o processamento e a análise dos dados.

## Objetivos Específicos:

- Realizar a orquestração de recursos em *cluster* de baixo custo;
- Comparar o desempenho de *clusters* bem 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 Contexto e Motivação
- 2 Justificativa
- 3 Objetivo
- 4 Revisão de literatura
  - Análise de dados
- 5 Método
- 6 Conclusão
- 7 Disponibilidade dos recursos deste trabalho

# Análise de dados

- Descisões em saúde costumam ser complexas – precisam de suporte científico (dados) e avaliação de Contexto
- Com o crescimento dos 3V's de dados na área da saúde (Big Data) processar e analisar esses dados tornou-se fundamental para tomada de decisões adequadas
- Desafios:
  - complexidade dos dados obtidos
  - ausência de validação de sistemas, métodos e ferramentas para o tratamento de dados na área
  - custos de novos equipamentos capazes de analisar tal volume
- Há grande oportunidade para a proposição de estratégias de processamento e análise de dados na área

- 1 Contexto e Motivação
- 2 Justificativa
- 3 Objetivo
- 4 Revisão de literatura
  - Alternativas *open source*
- 5 Método
- 6 Conclusão
- 7 Disponibilidade dos recursos deste trabalho



## Alternativas *open source*

- Considerando
  - O escopo deste trabalho
  - As estratégias para processamento e análise de dados disponíveis no mercado

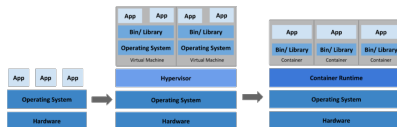
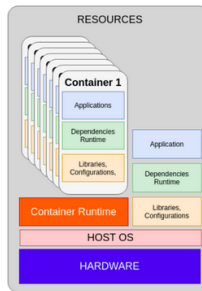
As soluções encontradas no mercado foram agrupadas em dois grupos:

- Soluções de Computação em nuvem privada:
  - ▶ Se estendem para além do propósito desse trabalho
  - ▶ Requisitos de hardware elevados
  - ▶ Complexidade de configuração devido a sua abrangência

# Alternativas *open source*

- Soluções de Orquestração de Containers:

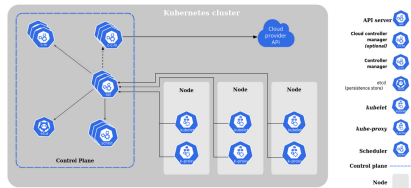
- Kubernetes®
- Apache Mesos®
- Hashicorp Nomad®
- Docker Swarm®



- 1 Contexto e Motivação
- 2 Justificativa
- 3 Objetivo
- 4 Revisão de literatura
  - Cluster orquestrador de container
- 5 Método
- 6 Conclusão
- 7 Disponibilidade dos recursos deste trabalho

# Cluster orquestrador de container

- Kubernetes®:
  - Origem de 15 anos de trabalho da Google (Borg)
  - Estrutura de objetos componentizados
    - ▶ Kube-apiserver
    - ▶ Kube-scheduler
    - ▶ Kube-controller-manager
    - ▶ Kubelet
    - ▶ Kube-proxy
    - ▶ Pod



1 Contexto e Motivação

2 Justificativa

3 Objetivo

4 Revisão de literatura

5 Método

- Abordagem

6 Conclusão

7 Disponibilidade dos recursos deste trabalho

# Abordagem I

Utilizar u *Cluster* ubernetes como plataforma de orquestração de cargas de trabalho em ambiente virtual.

- Cargas de trabalho:
  - Análise de tendência de uso de azitromicina entre 2014 e 2021
- Ambientes virtualizados (*Host* do cluster):
  - completa - *Hypervisor* tipo 2
  - sistema operacional - contêineres
- máquinas subutilizadas
- redução do CAPEX

# Abordagem

O uso de conceitos e metodologias de DevOps:

- CI (integração contínua)
- CD (entrega contínua)
- Monitoramento
  - método USE, parâmetros de utilização, saturação e erro
  - métricas definidas por parâmetro
  - relaciona desempenho dos nós virtuais.

1 Contexto e Motivação

2 Justificativa

3 Objetivo

4 Revisão de literatura

5 Método

- Especificações

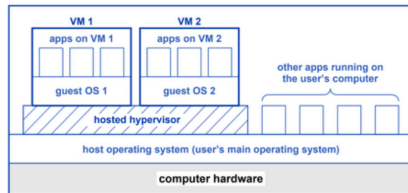
6 Conclusão

7 Disponibilidade dos recursos deste trabalho



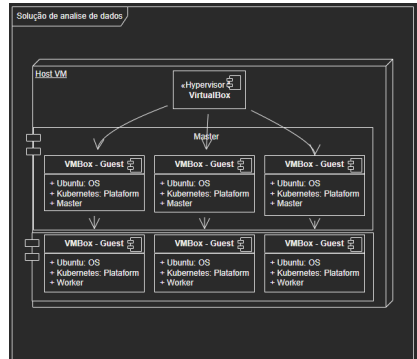
# Especificações I

- Cluster:
  - Virtualização:
    - ▶ Maquinas Virtuais (VMs) (*Hypervisor* tipo 2)
    - ▶ Contêineres Aninhados (Docker In Docker, ou DinD)
  - Ambiente Virtual:
    - ▶ arquitetura: **amd64**
    - ▶ 1 vCPU
    - ▶ 2 GB de RAM
    - ▶ 6-8 máquinas



## Especificações II

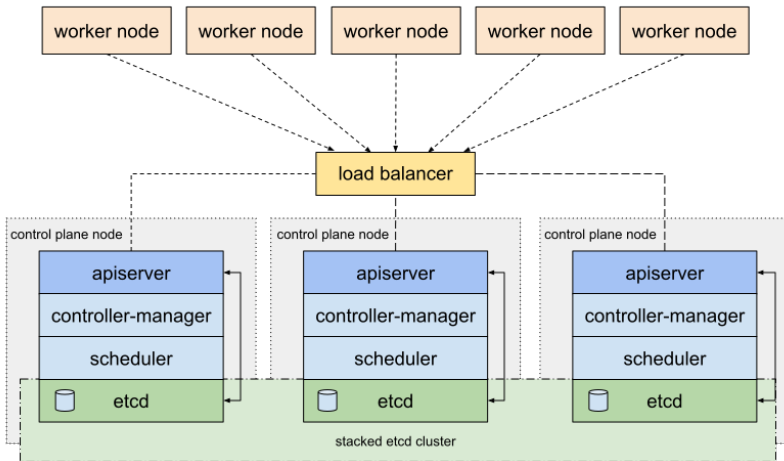
- Definição máquina *host*:
  - o *host* será um laptop, contendo configuração de 4vCPUs e 16GB de RAM
  - Hospedará as máquinas virtualizadas, pertencentes ao cluster
- Definição máquina *guest*:
  - como já descrito em 2 cenários: Virtualização completa, e em containers
  - Nós de *Cluster* Kubernetes (objeto de monitoramento)



- 1 Contexto e Motivação
- 2 Justificativa
- 3 Objetivo
- 4 Revisão de literatura
- 5 Método
  - Arquitetura Orquestrador
- 6 Conclusão
- 7 Disponibilidade dos recursos deste trabalho

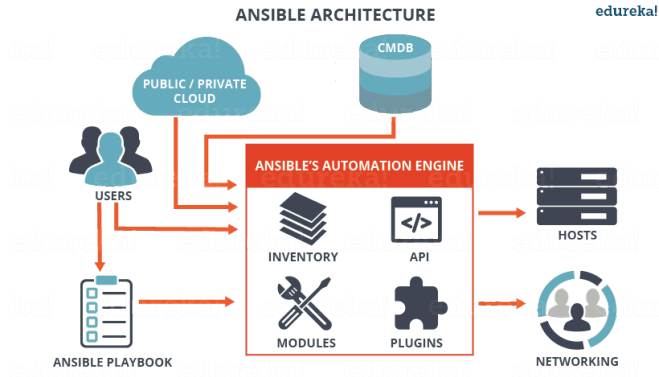
# Arquitetura Orquestrador

kubeadm HA topology - stacked etcd



- 1 Contexto e Motivação
- 2 Justificativa
- 3 Objetivo
- 4 Revisão de literatura
- 5 Método
  - Gerenciamento de configuração
- 6 Conclusão
- 7 Disponibilidade dos recursos deste trabalho

# Gerenciamento de configuração



1 Contexto e Motivação

2 Justificativa

3 Objetivo

4 Revisão de literatura

5 Método

- Monitoramento

6 Conclusão

7 Disponibilidade dos recursos deste trabalho

# 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



1 Contexto e Motivação

2 Justificativa

3 Objetivo

4 Revisão de literatura

5 Método

- Comparação entre tipos de virtualização

6 Conclusão

7 Disponibilidade dos recursos deste trabalho

# Comparação entre tipos de virtualização I

- macrobenchmark (system level benchmark) – Teste utilizando uma solução avaliando tempo de execução métricas de Desempenho (nós do cluster, *guests*):
- Taxa de Utilização de CPU e Memória
- Taxa de saturação de CPU e Memória
- Métricas de APM:
- Tempo Médio de todas as cargas de trabalho e variabilidade
- Método base utilizado para coleta de informações:
- Metodo USE de avaliação (Checklist Linux)

Caso base – comparação com processo de análise em *bare metal* 4vCPU, 16 GB de RAM – totalizando o poder computacional total do *cluster* proposto

1 Contexto e Motivação

2 Justificativa

3 Objetivo

4 Revisão de literatura

5 Método

- Análise de dados

6 Conclusão

7 Disponibilidade dos recursos deste trabalho

## Exemplo da 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

1 Contexto e Motivação

2 Justificativa

3 Objetivo

4 Revisão de literatura

5 Método

- Cronograma

6 Conclusão

7 Disponibilidade dos recursos deste trabalho

Fases de Projeto	Atividades do TCC	Ondas		Objetivos	Datas		Cascata
		Atividades			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 Delas Trabalho Inicial Monografia & Versão Final da Monografia	Especificação e critérios de aceitação			12	26/12/2021	08/01/2022
		Levantamento de Requisitos		Elaborar detalhamento da solução, mapear fronteiras da solução, identificar riscos ao projeto e propor desenho inicial da solução	13	09/01/2022	15/01/2022
		Levantamento de Lista de Materiais e softwares			15	16/01/2022	29/01/2022
		Apresentação dos 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			21	06/03/2022	12/03/2022
		Instrumentação (software) e verificação		Produção, Inspeção, Verificação e Validação da solução proposta.	26	13/03/2022	16/04/2022
		Implementação da análise e verificação			28	17/04/2022	30/04/2022
		Testes de Validação			30	01/05/2022	14/05/2022
Utilização & Suporte		Coleta dos resultados		Captação da utilização em cenário real em projeto de pesquisa parceiro	31	15/05/2022	21/05/2022
		Discussão dos resultados obtidos			33	22/05/2022	04/06/2022
Encerramento		Definição de próximas etapas		Estudo do caso de uso e sumarização dos resultados para apresentação da solução junto a banca	34	05/06/2022	11/06/2022
		Formalização dos trabalhos e apresentação			35	12/06/2022	18/06/2022

## Conclusão

- Entendimento da complexidade dos fatores considerados no processo de decisão em saúde
- Análise dos impactos sociais-econômicos relativos a restrição orçamentária na ciência
- Seleção de tecnologias com base em requisitos e restrições
- Desenho de uma estratégia de extração de informações em saúde
- Avaliação de diferentes tipos de virtualização e sua utilização

Trabalhos futuros contemplarão a implementação, testes e coletas de dados para avaliação comparativa das virtualizações propostas.

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.

# Disponibilidade dos recursos deste trabalho

- Github – Monorepo
- Pipeline

Todos os componentes definidos neste trabalho estarão contidos em um ou mais repositórios públicos, sob a licença pública geral GNU versão 3, para livre acesso.



# Referências I

OBRIGADO  
:)