UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

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

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Sumário

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

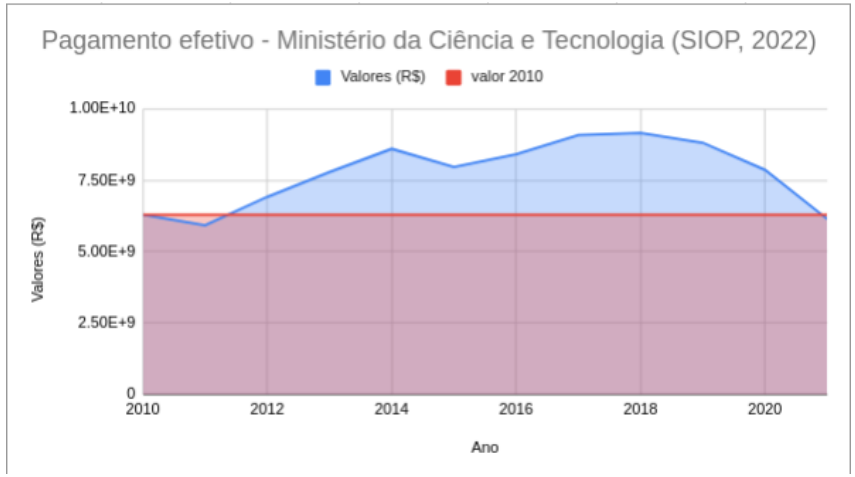
Introdução - Motivação

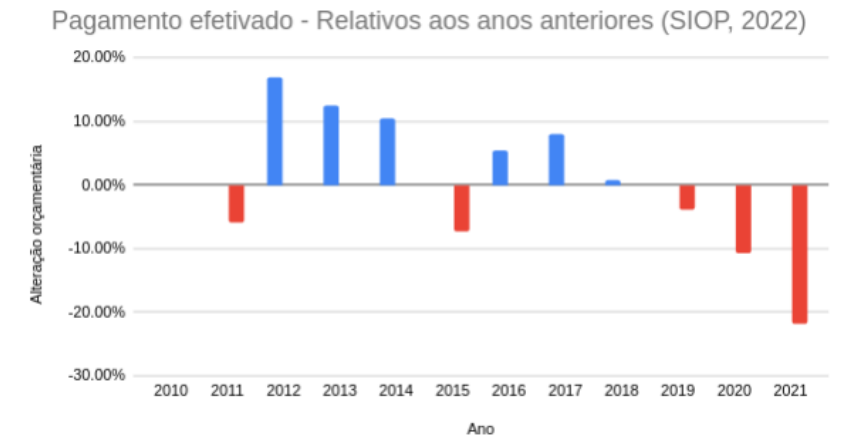
- Uso de ferramentas de analise dados em saúde
- Integração de Sistemas de Informação em Saúde e necessidade de facilitar processo de análise de grandes volumes de dados
- Disponibilidade de dados pelo Decreto nº 8.777 e a necessidade de extração de informações.

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Introdução - Justificativa

- Restrição Orçamentária
  - Diminuição de verbas para ciência e técnologia $-2,32\%$
  - Aumento do dólar em mais de $3,27\%$
- Disponibilidade de estratégias de analise





UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Introdução - Justificativa

- Restrição Orçamentária
  - Diminuição de verbas para ciência e técnologia $-2,32\%$
  - 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

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

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

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Introdução - Abordagem

- Aplicação de abordagem DevOps:
  - *Shift Right* - Fazes finais do *SDLC* (Ciclo de vida de)
  - CI (integração contínua) e CD (entrega contínua) - deploy da aplicação e inicio 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

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Objetivo

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 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;

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Revisão de literatura- Análise de dados

- Complexidade de tomar descisão em saúde
- Definição de Big Data 5 Vs, complexidade e Destruturação
- Complexidade de relacionar dados por multifatoriedade
- Não consolidação de métodos de uso de Big Data em saúde, especialmente em estudos quantitativos, foco em custo e decisões clínicas

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

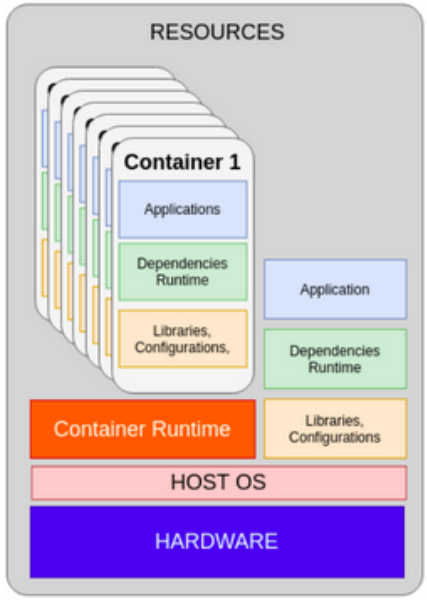
Revisão de literatura - Alternativas open source

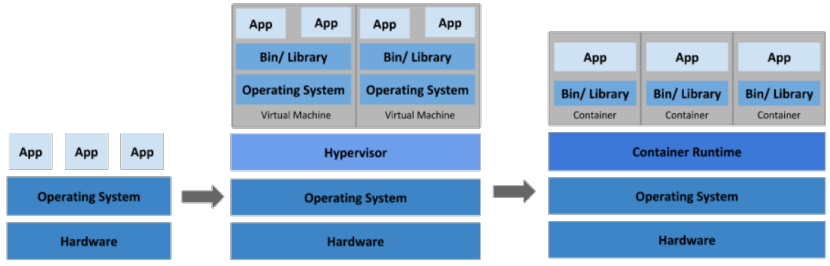
- Agrupamento em categorias:
  - Computação em nuvem privada:
    - Tecnologias: OpenStack®, CloudStack®
    - Requisitos exigentes
    - Complexidade: SaaS, PaaS e SaaS

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Revisão de literatura- Alternativas open source

- Orquestração de Containers:
  - Kubernetes®
  - Apache Mesos®
  - Hashicorp Nomad®
  - Docker Swarm®

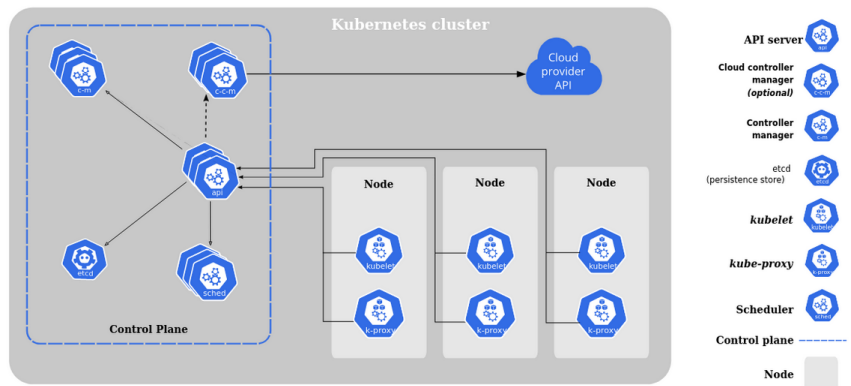




UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Revisão de literatura- 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



UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

1 Introdução

2 Objetivo

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

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

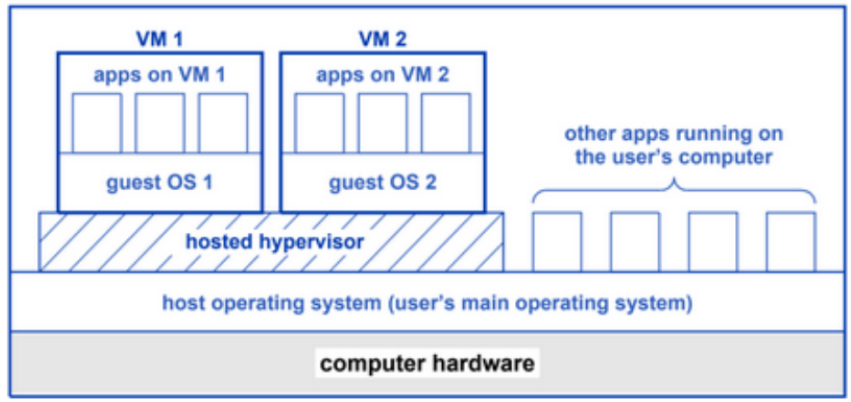
Método - Disponibilidade dos recursos deste trabalho

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 https://github.com/felipefrocha/esufmg-tcc Repositório

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Método - Especificação dos nós integrantes cluster de baixo custo

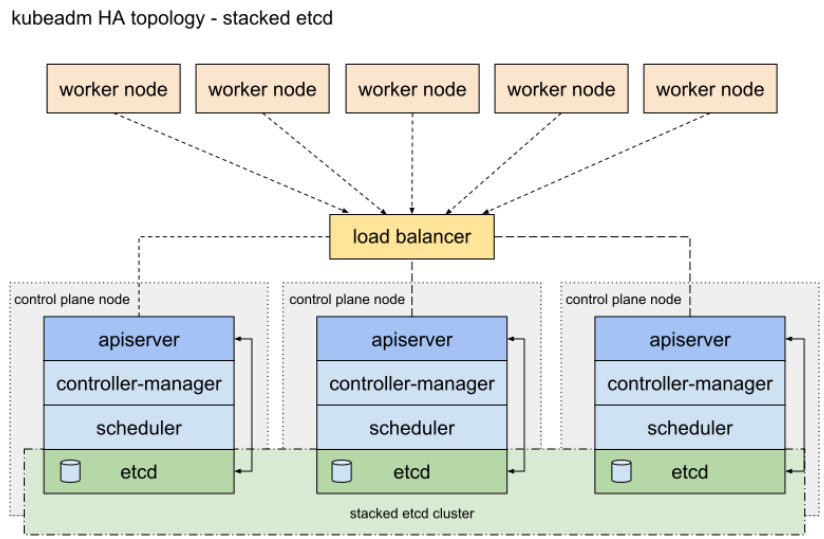
- Cluster Simulado:
  - Virtualização:
    - Maquinas Virtuais (VMs) (*Hypervisor* tipo 2)
    - Contêineres Aninhados (Docker In Docker, ou DinD)
  - Especificações de hardware 1vCPU, 2 GB de RAM;
- provisionamento em 2 etapas
- máquinas subutilizadas
- CAPEX



UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Método - Plataforma de orquestração de carga de trabalho

- Multi-master Etcd atachado
- Arquitetura sugerida para produção
- Alta disponibilidade do cluster
- Recursos limitados



UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Método - Plataforma de orquestração de carga de trabalho

- Implantação da carga de Trabalho
  - Container
  - Parametrizável
  - Volume compartilhado
- Ciclo de vida da aplicação
  - Monorepo
  - Padronização de código
  - Sincronização
  - Deploy em Pipeline

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Método - Configuração e provisionamento do cluster

- Gerenciador de Congiração (Ansible®)
- *Agentless*
- Idempotência
- Gerenciamento de inventário
- SSH - Escolha do algoritimo de criptografia

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

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

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

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

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Método - Comparação entre tipos de virtualização

- A arquitetura x86 comum e maior poder computacional
- macrobenchmark (system level benchmark)
- Parametros de tempo, taxa de utilização de memoria e processamento
- Maquina hospedeira e virtuais serão avaliadas durante o processamento
- Metodo USE de avaliação
- APM Application Performance Management) associada a aplicação da carga de trabalho por *OpenTelemetry*

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Ondas				Dates			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 revisão bibliográfica		3	31/10/2021	06/11/2021	
Concepção	Visão Geral do Projeto	Descrição formal dos *stakeholders*	Identificar publico 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 da Defesa	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	
	Texto Inicial Monografia & Versão Final da Monografia	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	
		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	
		Discussao dos resultados obtidos.		33	22/05/2022	04/06/2022	
Encerramento		Definição de proximas etápas	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 trabalho e apresentação		35	12/06/2022	18/06/2022	

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Conclusão

- Avaliação de diferentes tipos de virtualização
- Seleção de plataforma de orquestração de cargas de trabalhos com base em requisitos e restrições
- Analise 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 descisão 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 maquina base.

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Referências I

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

OBRIGADO
:)