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

Sumário

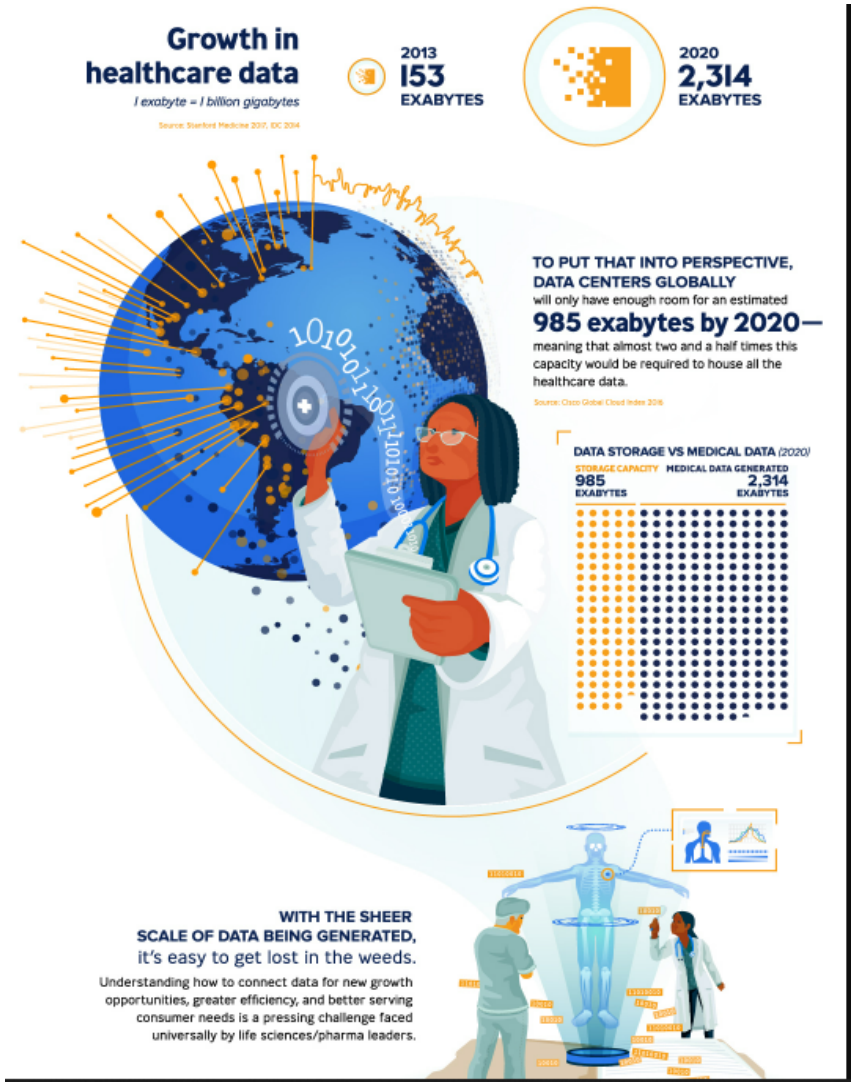
UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Contexto e Motivação I

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

Várias ferramentas estão disponíveis para Transforma-los em informações e embasar descisões

Contexto e Motivação II



Isso também acontece na área da saúde

Porém o uso dessas ferramentas nessa área, para transformar dados em informação, ainda é pouco significativo

Contexto e Motivação III

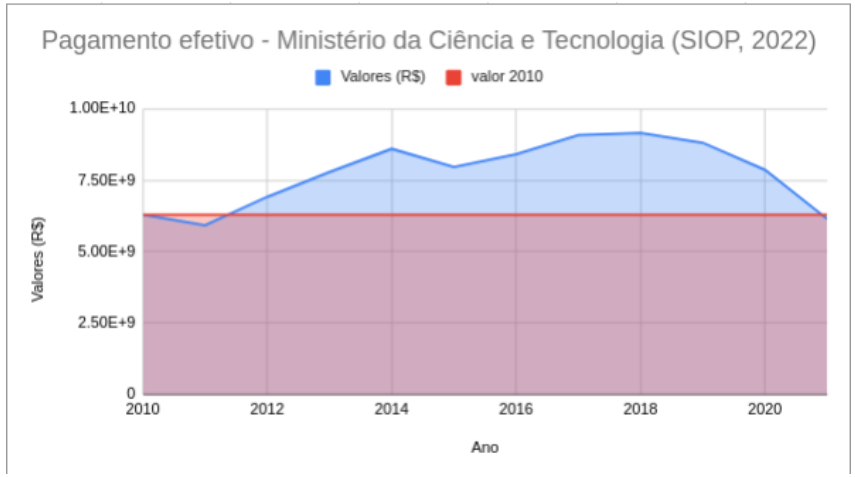
- : Tendência crescente de trabalho interdisciplinar
- : Potencial de melhora do sistema de Saúde através de análise de dados
- : Necessário Propor e Validar estratégias que sejam viáveis e facilitem o processamento de analise de grande volume de dados produzido na área

Contexto e Motivação IV

- : No Brasil, dados do Sistema de Informação em Sáude (SIS) são disponibilizados desde 2016
- : **Faltam recursos** e Estratégias Viáveis para essa elboração.

Introdução - Justificativa

- Restrição Orçamentária
  - Diminuição de verbas para ciência e técnologia $-2,32\%$, mesmo com o aumento de base de alunos
  - Aumento do dólar em mais de $3,27\%$ diminuindo o poder de compra
  - Tomada de decisão em saúde, mais de 152mi de Brasileiros dependem exclusivamente do SUS
- Necessidade de dispor estratégias de análise de dados



Introdução - Justificativa

- Restrição Orçamentária
  - Diminuição de verbas para ciência e técnologia $-2,32\%$, mesmo com o aumento de base de alunos
  - Aumento do dólar em mais de $3,27\%$ diminuindo o poder de compra
  - Tomada de decisão em saúde, mais de 152mi de Brasileiros dependem exclusivamente do SUS
- Necessidade de dispor estratégias de análise de dados

Introdução - Abordagem

UFMG
UNIVERSIDADE FEDERAL DE MINAS GERAIS

Utilizar um Cluster Kubernetes como plataforma de orquestração de cargas de trabalho em ambiente virtual.

- Cargas de trabalho:
  - Analise de tendencia de uso de azitromicina entre 2014 e 2021
- Ambientes virtualizados (simulando *Host* do cluster):
  - completa - *Hypervisor* tipo 2
  - sistema operacional - contêineres
- Ambiente virtual:

  Simulação de máquina de baixo poder computacional:
  - 1 vCPU
  - 2 GB de RAM
  - 6-8 máquinas

Essa Abordagem visa comparar o desempenho desses ambientes simulados, e validar o uso de computadores de baixo poder computacional, no processo de análise de dados de grande volume

Introdução - Abordagem

O uso de conceitos, métodose o uso de ferramentas complementares na aplicação da cultura DevOps em ambientes produtivos, permitirá o deployment simplificado melhorando a agilidade e diminuindo a complexidade e operação/sustentação do cluster

- Conceitos como:
  - CI (integração contínua)
  - CD (entrega contínua)
- Uso do metodo USE (utilização, saturação e erro). Esse método propoe um checklist de métricas a serem coletadas e a avaliação de três paramêtros por meio dessas métricas, relacioando assim o performance da carga de trabalho (aplicação) e o desempenho dos nós do cluster sob monitoramento.

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Objetivo

UFMG
UNIVERSIDADE FEDERAL DE MINAS GERAIS

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

Revisão de literatura- Análise de dados

UFMG
UNIVERSIDADE FEDERAL DE MINAS GERAIS

- 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 tornouse fundamental para tomada de descisões adequadas
- Desafios:
  - complexidade dos dados obtidos
  - ausencia 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 anális de dados na área

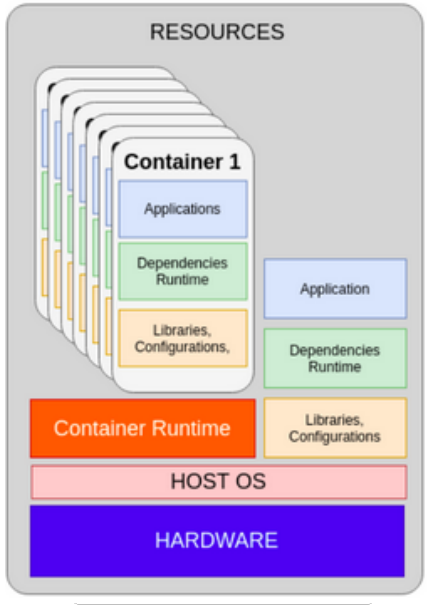
Revisão de literatura - Alternativas open source

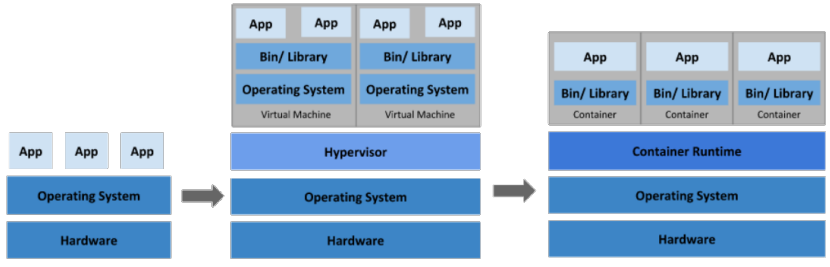
- Considerando
  - O escopo deste trabalho
  - As estratégias para processamento e análise de dados disponíveis no mercado

  As soluções encontradas no mercado foram agrupadas em dois grupos:
  - Soluções de Computação em nuvem privada:
    - Se extendem para além do proposito desse trabalho
    - Requisitos de hardware elevados
    - Complexidade de configuração devido a sua abrangência

Revisão de literatura- Alternativas open source

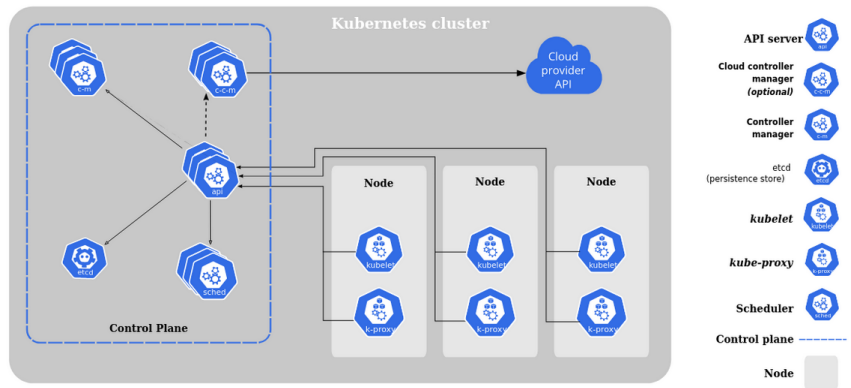
- Soluções de Orquestração de Containers:
  - Kubernetes®
  - Apache Mesos®
  - Hashicorp Nomad®
  - Docker Swarm®





Revisão de literatura- Cluster orquestrador de container

- Kubernetes®:
  - Origem de 15 anos de trabalho da Google (Borg)
  - 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

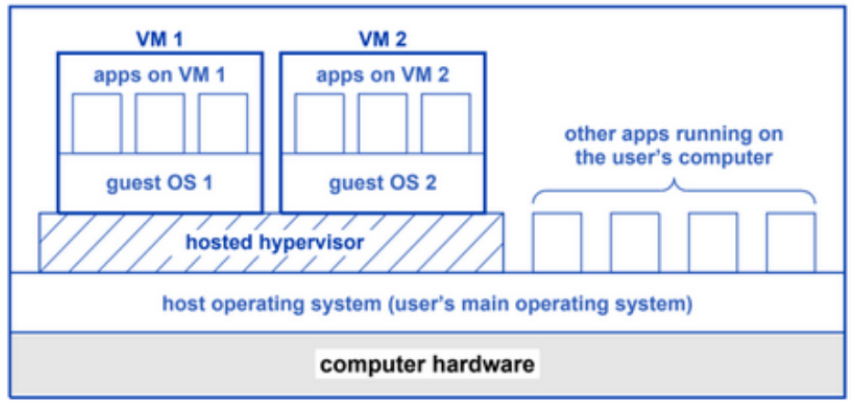
- Especificação dos nós integrantes cluster de baixo custo
- Plataforma de orquestração de carga de trabalho
- Configuração e provisionamento do cluster
- Análise de dados
- Monitoramento
- Comparação entre tipos de virtualização
- Cronograma

5 Conclusão

6 Disponibilidade dos recursos deste trabalho

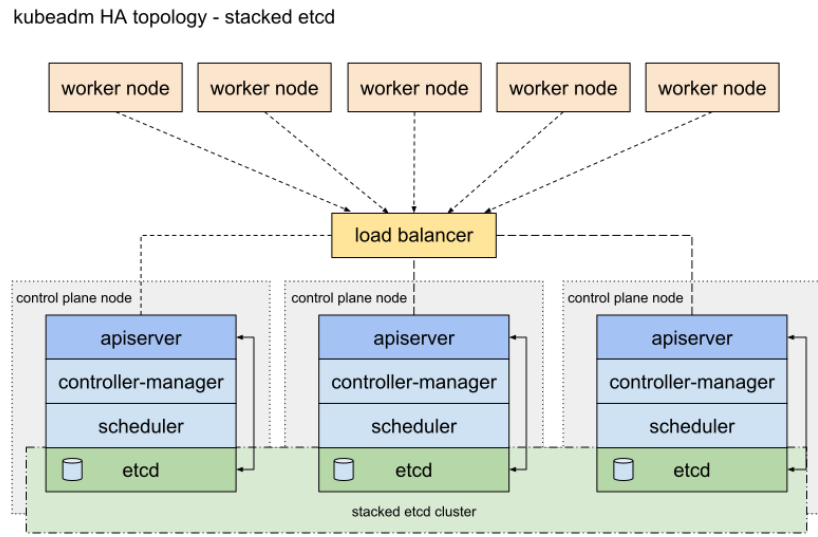
Método - Especificação dos nós integrantes cluster de baixo custo

- 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

- Arquitetura sugerida para produção:
  - Multi-master com Etcd junto ao nó master
- Alta disponibilidade do cluster
- Recursos de hardware limitados



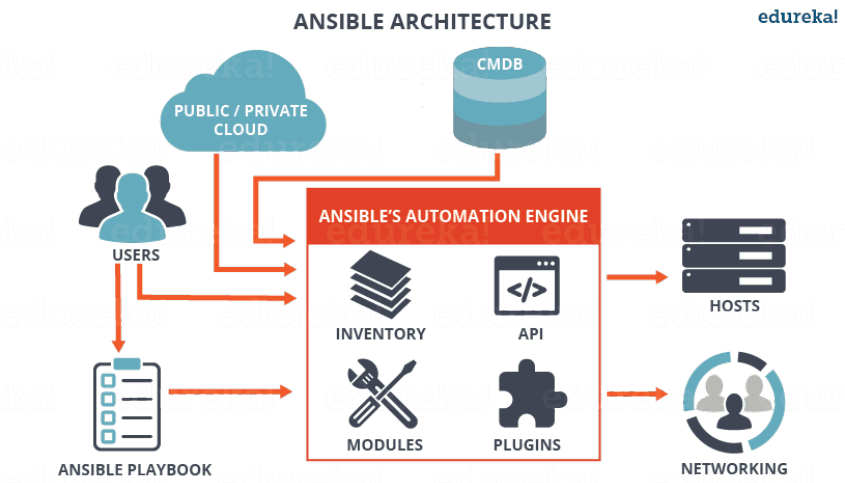
Método - Plataforma de orquestração de carga de trabalho

- Implantação da carga de Trabalho
  - Container
  - Parametrizável
  - Volume compartilhado

Método - Configuração e provisionamento do cluster

O uso de gereniadores de configuração garantem o versionamento das configurações permitindo maio controle sobre as propriedades dos *assets* gerenciados (Ansible®)

- *Agentless*
- Idempotência
- Gerenciamento de inventário
- SSH - Escolha do algoritimo de criptografia



Método – Análise de dados

UFMG
UNIVERSIDADE FEDERAL DE MINAS GERAIS

- "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

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

- *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
- 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

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

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

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

OBRIGADO
:)