UNIVERSIDADE FEDERAL DE MINAS GERAIS
CURSO DE ENGENHARIA DE SISTEMAS

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 Gerias

BELO HORIZONTE
JANEIRO DE 2022

FELIPE FONSECA ROCHA

ANÁLISE DE DADOS UTILIZANDO CLUSTER DE BAIXO CUSTO: COMPARAÇÃO DE DESEMPENHO DE AMBIENTES VIRTUAIS

Trabalho de Conclusão de Curso I apresentado ao Curso de graduação em Engenharia de Sistemas do Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do título de Bacharel em Engenharia de Sistemas.

Orientador: Ítalo Fernando Scotá Cunha
Universidade Federal de Minas Gerias

UNIVERSIDADE FEDERAL DE MINAS GERAIS
CURSO DE ENGENHARIA DE SISTEMAS
BELO HORIZONTE
JANEIRO DE 2022

Dedico este trabalho àquela que esteve comigo na jornada dessa graduação, minha esposa. Dedico também aos meus avós, pelo suporte aos meus estudos e modelos de cidadãos e chefes de família, mas que não puderam ver a conclusão. E também aos meus pais que me incentivaram a nunca desistir.

Agradecimentos

Agradeço a minha esposa, que em seu apoio, não só me ensinou o valor da acadêmia, mas me mostrou o papel da universidade no processo de transformação da sociedade na ostentação da tríade: ensino, pesquisa e extensão; sem nunca privilegiar um em detrimento do outro.

Ao meu orientador, Prof. Dr. Ítalo Fernando Scotá Cunha pela ajuda durante o processo de formulação de um trabalho válido e suas contribuições a este.

Aos docentes exemplares do curso de Engenharia de Sistemas, que em muito contribuíram para minha formação.

À Prof. Dra. Ana Liddy Cenni de Castro Magalhães, por sua excelente coordenação e o secretário do Colegiado de Engenharia de Sistemas, Júlio César Pereira de Carvalho, por sua disponibilidade e suporte.

Ao Centro de Estudos de Medicamentos da Faculdade de Farmácia da UFMG e todos os seus membros, por sua iluminação sobre o que é a extensão aliada ao ensino e orientada a pesquisa.

Aos colegas do curso que dividiram seus conhecimento e experiências.

A minha família, pela força motriz de tudo o que construo.

Aos meus amigos, dentre eles graduados, mestres e doutores pela sua importância ao longo da minha trajetória.

”Os livro têm muita coisa escrita, é o que ele diz

Em pleno berço de Machado de Assis, olha o que eu sou obrigado a ouvir

Nega a ciência, esconde a doença, só negligência

Essa é a sentença pra falta de consciência que colocou ele ali”

(Chico César, DJ Caique, Rashid, Diário de Bordo 6)

Resumo

Análise de grandes massas de dados, por vezes requer um, ou mais computadores de alto desempenho para tornar viável a extração de informações. Na área da pesquisa, principalmente de instituições publicas, há uma restrição orçamentaria que nos últimos anos apresentou uma diminuição do valor disponibilizado. Dessa forma, áreas de estudo, como a da saúde, que possuem um volume de dados considerável, vêm enfrentando problemas na aquisição desses equipamentos que habilitam essas análises. Uma alternativa a necessidade de computadores de auto desempenho é o uso de técnicas de computação distribuída. O objetivo deste trabalho visa realizar o estudo das ferramentas disponíveis no mercado e propor uma solução de implementação e comparação, para no trabalho posterior, comparar o desempenho de orquestração de recursos em *cluster* de baixo custo em ambientes com virtualização completa e virtualização baseada em contêineres para o processamento e a análise de dados em saúde, como tendência de consumo de azitromicina entre os anos 2014 e 2021. Ainda nesse trabalho apresenta-se uma avaliação de impactos econômicos na processamento de dados e aspectos sociais de contribuição resultantes dessa pesquisa.

Palavras-chave: Kubernetes®. Virtualização. Contêineres. Hypervisor Tipo 2. Análise de dados.

Abstract

Analysis of large datasets sometimes requires one or more high-performance computers to make extracting information feasible. In the area of research, mainly in public institutions, there is a budget constraint that in recent years has shown a decrease in the amount available. Thus, areas of study, such as health, which have a considerable volume of data, have been facing problems in acquiring the equipment that enables these analyses. An alternative to the need for self-performing computers is the use of distributed computing techniques. The objective of this work is to carry out a study of the tools available on the market and propose an implementation and comparison solution, in order to compare the performance of low-cost resource orchestration in *cluster* in environments with full virtualization and virtualization. based on containers for the processing and analysis of health data, such as azithromycin consumption trend between the years 2014 and 2021. This work also presents an assessment of economic impacts on data processing and social aspects of contribution resulting from this research.

Keywords: Kubernetes. Virtualization. Containers. Hypervisor Type 2. Data analysis.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

1.1 Motivação

No contexto da análise de dados, diferentes ferramentas estão disponíveis para transformá-los em informação, contudo o uso dessas ferramentas na área da saúde ainda é pouco significativo (GALVãO; VALENTIM, 2019). Frente a uma tendência crescente de interconexão entre diferentes áreas do conhecimento e do potencial da análise de dados possibilita para melhoraria do sistema de saúde, se faz necessário propor e validar estratégias que permitam o avanço na integração de dados entre diferentes Sistemas de Informação em Saúde (SIS), e que facilitem o processamento e análise do grande volume de dados produzidos e disponibilizados nesses sistemas (GALVãO; VALENTIM, 2019; MEHTA; PANDIT, 2018a).

Atualmente, conforme determinação do Decreto nº 8.777, de 11 de maio de 2016, que instituiu a Política de Dados Abertos do Poder Executivo Federal (BRASIL, 2016), diversos dados dos SIS são disponibilizados de forma pública. No entanto, apenas a disponibilização dos dados em si não garante que os mesmos poderão ser analisados e com isso produzir informação relevante para as políticas públicas na área da saúde.

1.2 Justificativa

Contudo, como observado nos últimos 10 anos (Tabela 1), a disponibilidade recursos financeiros efetivos para ciência e tecnologia no Brasil têm oscilado. Nos anos de 2018 a 2021 sofreu reduções acentuadas, o que torna o acesso a recursos que viabilizem a realização de análise dos dados em ferramentas e infra estruturas tradicionais ou ainda a proposta de novos, limitados.

Tomando como exemplo os bancos de “Vendas de Medicamentos Controlados e Antimicrobianos - Medicamentos Industrializados”, objeto deste trabalho, estão disponíveis cerca de 70 GB e com mais de 500 milhões de linhas de dados sobre a comercialização de medicamentos no país. Logo, tão importante quanto a disponibilidade pública dos dados é fundamental encontrar estratégias técnicas e economicamente viáveis a fim de possibilitar que pesquisadores em todo o país possam contribuir com a análise e a interpretação desses dados, mesmo frente a baixa disponibilidade de recursos financeiros e de infraestrutura, como servidores de alta performance (HPC), por exemplo.

Tabela 1 – Pagamento efetivo - Ministério da Ciência e Tecnologia

Ano	**Objetivo**
2010	R$ 6.288.931.123,00
2011	R$ 5.918.584.706,00
2012	R$ 6.918.288.201,00
2013	R$ 7.787.464.592,00
2014	R$ 8.598.785.224,00
2015	R$ 7.964.319.815,00
2016	R$ 8.404.014.691,00
2017	R$ 9.085.620.227,00
2018	R$ 9.157.748.260,00
2019	R$ 8.812.096.752,00
2020	R$ 7.859.851.948,00
2021	R$ 6.142.873.884,00

Fonte: SIOP consulta realizada em Janeiro de 2022

1.3 Objetivos

Diante disso, com a realização desse trabalho espera-se oferecer uma alternativa para análise de grandes volumes de dados que possua baixo custo financeiro, menor complexidade de configuração, maior efetividade (menor tempo de análise) e que não seja dependente da disponibilidade ou uso de recursos dedicados, como HPC, às análises, como é o caso de outras alternativas open source atualmente disponíveis - ex.: OpenStack, CloudStack etc.

Deste modo, espera-se demonstrar comparativamente a implementação de uma solução para análise de dados em plataformas de orquestração de containers, que permita recrutar computadores comuns para essa análise. E assim, espera-se, superar de maneira custo-efetiva um problema de restrição orçamentária e técnica para instituições públicas e grupos de pesquisa que realizam análises de grande volumes de dados, no caso desse trabalho para área da saúde, utilizando uma tecnologia já amplamente empregada no setor privado. O que viabiliza o suporte de estudantes e/ou profissionais das áreas de Engenharias e Computação. Espera-se ainda, contribuir para que os dados públicos em saúde sejam analisados com maior frequência e menor restrição, gerando indicadores melhores e atualizados para melhor tomada de decisão em saúde.

1.4 Definição e abordagem

A proposta do trabalho visa comparar a utilização de cluster de Kubernetes®como plataforma de orquestração de cargas de trabalho em dois tipos ambientes virtualizados, utilizando como carga de trabalho a análise de tendência de consumo de azitromicina no Brasil entre os anos de 2014 e 2021. Tendo como principal resultado uma análise

comparativa de desempenho dos ambientes e uma proposta de utilização dessa plataforma em computadores do tipo desktop como alternativa a HPC. A utilização da plataforma visa validar seu uso para orquestração de tarefas em paralelo, durante a análise permitindo o uso de diversas máquinas. Para esse trabalho a comparação será em virtuais (VMs) e containers aninhados, no caso deste trabalho 6, com capacidades de processamento semelhantes a computadores desktop de 2 GB (Gigabytes) de RAM (Random Access Memory) e 1 vCPU (virtual Central Process Unit). Essa restrição para ambiente virtualizado será realizada por configuração de API do hypervisor ou por restrição de cgroups. Permitindo que a análise de grandes massas de dados (maiores que 50 GB) possam ser feitas sem o uso de HPC.

A abordagem de DevOps (BASS et al, 2015) para tornar o provisionamento, integração e deploy da infra estrutura, bem como os componentes de análise desse utilizados neste trabalho incluem o conceito de CI (continuous integration), CD (continuous delivery) IaC (Infrastructure as Code) visa tornar a configuração e disponibilização desse cluster mais ágil, diminuindo assim a necessidade de operação e também de manutenção do mesmo.

Para a análise de dados, utilizando a estratégia descrita, propõe-se analisar as tendências de consumo da azitromicina no período de 2014 a 2021, essa análise é objeto de carga de trabalho a ser orquestrado de maneira distribuída no cluster para validação de seu desempenho nos ambientes propostos.

O trabalho não foca na realização de interpretação da informação gerada pelo banco, garantindo assim apenas o resultado correto da análise citada como carga de trabalho para comparação. Também não está sendo proposta uma metodologia de análise do banco referenciado. Mas a avaliação das tecnologias empregadas para orquestração das tarefas, comparação de desempenho entre as alternativas da implementação da plataforma e sua implementação como proposta para uso mais amplo nas instituições sob restrição orçamentária, com o fim de continuar a realizar analises de dados, ainda que sem hardware adequado.

1.5 Organização do trabalho

Esse trabalho irá apresentar a fundamentação teórica e revisão de literatura na seção 2, apresentando as leituras que embasaram toda a construção do projeto. Na seção 3 apresenta-se a metodologia utilizada para a construção dos componentes do projeto e a forma de avaliação de desempenho dos ambientes propostos bem como a forma de avaliação.

2 Fundamentação Teórica

2.1 Análise de dados em saúde

Tomar decisões em saúde sempre precisam do suporte de um profissional especializado na temática a ser decidida. Uma série de técnicas são aplicadas por esse profissional para avaliação do contexto, uma vez que os dados puros não são suficientes pela multifatoriedade (ANDRADE, 2008; RESENDE; VIANA; VIDGAL, 2009). Com o grande número de sistemas de auxílio em saúde, sejam de ordem regulatória ou ainda por iniciativas privadas, o volume de dados cresce exponencialmente, produzindo fenômeno de Big Data. Uma definição conhecida desse conceito aponta volume, variedade e velocidade como os três vetores relacionados à produção massiva de dados (LANEY et al., 2001). Outras características foram adicionadas a essa primeira definição de Big Data, tais como veracidade (SCHROECK et al., 2012), complexidade e desestruturação (CORPORATION, 2012), valor (DIJCKS, 2013).

A multifatoriedade citada acima é um aspecto que se relaciona a complexidade dos dados em grande volumes, é cada vez mais difícil estabelecer uma relação de causa efeito para a tomada de decisão segura, sem auxílio ou sumarização desses dados em informações mais tangíveis e associadas. Ao mesmo tempo que é extremamente difícil de descrever um algoritmo suficiente que permita analisar todos os dados de contexto e relacioná-los de forma que possa ser utilizado como substituição ao conhecimento tácito de um profissional da saúde experiente (FACELI et al., 2011).

O suporte de sistemas, métodos e ṕraticas que suportem tanto a assistência em saúde no tratamento de Big Data ainda não é suficientemente consolidado. Faltam evidências de aplicações especialmente quantitativas que validem seu uso nesse campo do conhecimento. Além disso boa parte das tecnologias relacionadas em diversos estudos, utilizam tecnologias que demandam recursos específicos. Não obstante essa análise, estão focadas na avaliação de dados para otimização de recursos, suporte a decisões clínicas e redução de custo do cuidado (MEHTA; PANDIT, 2018b).

Havendo ainda uma grande oportunidade de estudo para a área de processamento de dados que sumarizam informações e tornem mais fácil seu estudo, correlação e posteriormente associação com demais bancos, ou conjuntos de informações que é justamente o que esse trabalho se propõe, ainda que o volume de dados da base analisada neste trabalho e sua composição não qualifiquem como Big data, sob a perspectiva de LANEY. Pode ser utilizado de base para uma nova forma de utilização de recursos computacionais na análise de grandes massas de dados e a posteriori Big Data.

2.2 Alternativas open source

A pesquisa sobre estratégias para a análise de dados retornou um grande número de artigos, sendo encontradas estratégias diversas descritas na literatura científica. Para fins de apresentação dessas soluções neste trabalho, elas foram agrupadas conforme níveis de complexidade de implementação, sempre mantendo o custo como restrição primária e o contexto de aplicação em instituições com grande restrição orçamentária. Diante disso, e tendo em vista o tamanho da base de dados de origem, durante o levantamento do projeto foram avaliadas soluções em duas categorias:

- Computação em nuvem privada; e
- Orquestração de Containers.

Na primeira categoria, a utilização de ferramentas como OpenStack®e CloudStack®foram consideradas. No entanto, observa-se que nesse tipo de utilização existe um overhead substancial, tanto em termos de complexidade de configuração, quanto em termos de hardware necessário para operar de maneira eficiente. Essas alternativas requerem maior capacidade computacional, conforme suas configurações recomendadas (CLOUDSTACK, 2022; OPENSTACK, 2022). Além disso, as soluções de nuvem privada estendem muito o propósito de orquestração de cargas de trabalho, provendo todo o conceito de infraestrutura como serviço (Iaas). E, dependendo da implementação e dos componentes utilizados, elas provêm plataforma como serviço (PaaS), que são categorias de abstração do hardware, configurações e sistemas de suporte/apoio como Sistemas Operacionais (OS) para aplicação (OPENSTACK, 2022; CLOUDSTACK, 2022; MELL; GRANCE, 2011). Logo, embora sejam alternativas populares, tornam o processo substancialmente mais complexo e, portanto, foram descartadas para essa avaliação tendo em vista o escopo deste trabalho.

Na segunda categoria, sistemas de orquestração de containers possuem um baixo overhead devido a arquitetura do container (Figura 1), compartilhando parte do kernel space. Logo, não necessita de uma camada de virtualização do sistema operacional, presente em virtualizações completas e gerenciadas por hypervisors (Figura)2). Vale ressaltar que essa solução oferece ainda algumas possibilidades como, por exemplo, o gerenciamento de capacidade dos nós do cluster para agendamento de tarefas. Dentre as plataformas disponíveis, o Kubernetes®é apontado como uma das principais soluções de orquestração de containers, tanto pela disponibilidade de features, quando pelos projetos em operação e pelo tamanho de sua comunidade (TRUYEN et al., 2021). O Kubernetes®tem o apoio de entidades como Cloud Native Computing Foundation (CNCF), que apoiam e supervisionam a plataforma de software - definida como peça importante de software, sob o qual diversos programas de aplicativos menores podem ser projetados para serem executados (PLATFORM. . . , 2022). Esse apoio tem como objetivo a expansão das capacidades, endereçando problemas conhecidos e situações de uso, bem como estabelecendo padrões para tecnologias que pertencem ao ecossistema de orquestração de containers.

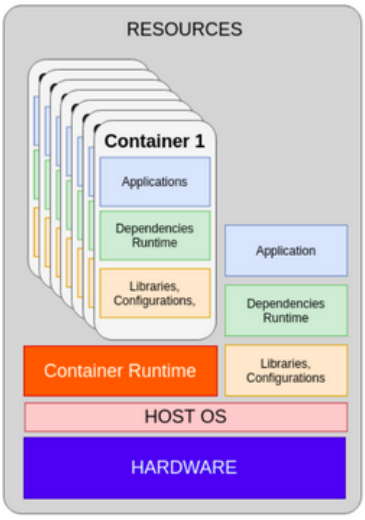

Figura 1 – Estrutura do container (elaborada pelo autor)

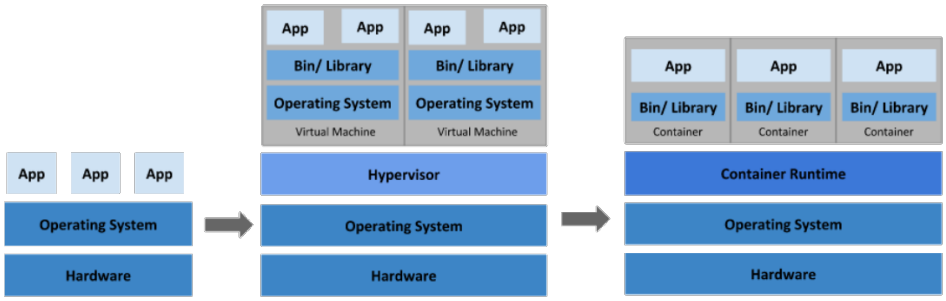

Figura 2 – Eras de deployments e sua evolução por tecnologia base (Fonte: The Linux Foundation®, 2021)

Soluções como Apache Mesos, Hashicorp Nomad, e Docker Swarm também foram avaliadas pelo estudo (TRUYEN et al., 2021), mas em todos os casos foram citadas diferenças significativas, especialmente no uso, sendo o Kubernetes®a melhor avaliada.

2.3 Cluster orquestrador de container

Kubernetes®é a consolidação de quinze anos de trabalho da Google®com orquestração de cargas de trabalho, processamentos *batch*, e um sistema interno de gerenciamento de cluster orientado a containers, o Borg (VERMA et al., 2015).

As estruturas básicas do Kubernetes®são divididas em componentes com atribuições bem definidas, como na Figura 3. Os componentes que são essenciais a proposta deste trabalho são:

- Kube-apiserver que concentra toda a api do kubernetes
- Kube-scheduler que avalia novos pods e em qual nó worker do cluster os mesmo

serão alocados

- Kube-controller-manager que comporta os objetos de controle do kubernetes
- Kubelet responsável por repassar comando do control plane, para o worker e comunicar com container runtime
- Kube-proxy responsável por toda a estrutura de redes nativa do kubernetes e pods do cluster
- Pod menor unidade de deployment podendo conter um conjunto de containers que compõem uma solução.

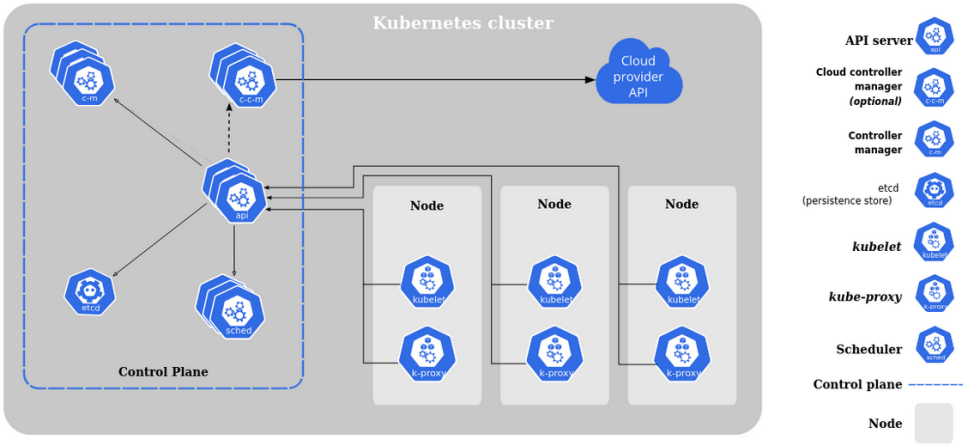

Figura 3 – Kubernetes arquitetura de componentes (Fonte: The Linux Foundation®, 2021)

3 Metodologia

3.1 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,).

3.2 Especificação dos nós integrantes cluster de baixo custo

A formação do cluster será de máquinas virtuais (VMs) e/ou físicas que possuem custo e capacidade computacionais mais baixos, com configurações comumente encontradas em computadores do tipo desktops, como os utilizados em residências, escritórios e laboratórios de informática genéricos.

A primeira versão dessa solução será realizada de maneira simulada, utilizando ambiente virtualizado com VMs que por sua vez serão provisionadas em hypervisors do tipo 2 (COMER, 2021) (*hosted hypervisor*), para garantir a simplicidade da implementação da primeira fase (TCC 1) e validar o conceito de provisionamento, configuração e *deploy* de aplicações testes, bem como laboratório das ferramentas de monitoramento que serão utilizadas, a serem descritas a seguir.

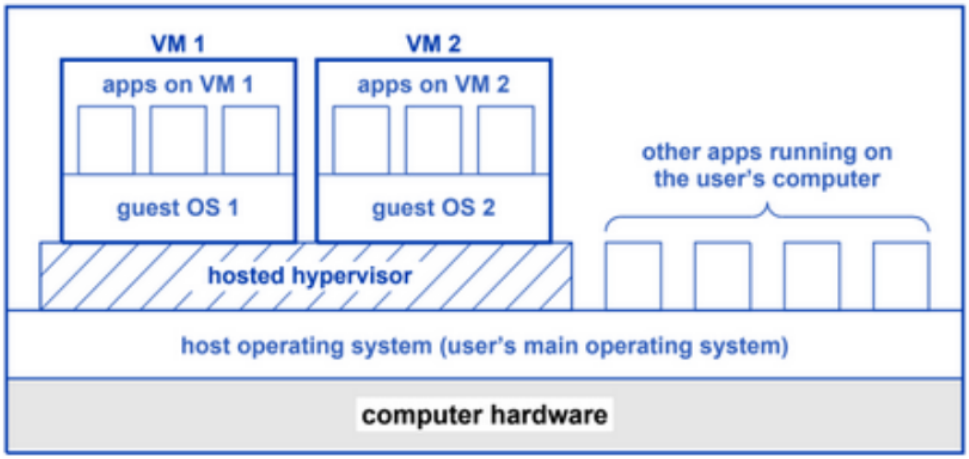

Figura 4 – Hosted Hypervisor diagrama representativo de componentes (Fonte: Google Images, 2021)

A versão final do projeto (ambas no TCC II) visa o provisionamento do cluster nos ambientes de teste configurados como descrito anteriormente, sendo que algumas

premissas serão utilizadas, como o uso de mesmo sistema operacional e versão do mesmo em cada um desses computadores. Garantindo a redução de configuração necessária para a implementação. Outra premissa utilizada para esse estudo, será a utilização de uma rede comum aos computadores.

O aproveitamento de computadores comuns já existentes e subutilizados seja por uso abaixo de sua capacidade ou ainda intervalos de ociosidade, qualifica o baixo custo da formação do cluster em questão, apresentando CAPEX (capital expenditure), ou investimento inicial mínimo, se não zero, necessitando da interdisciplinaridade sugerida e defendida dentro das instituições para o qual esse trabalho intenta apresentar uma alternativa.

3.3 Plataforma de orquestração de carga de trabalho

A plataforma de cluster e orquestração de cargas de trabalho utilizadas nesse trabalho será o Kubernetes. A arquitetura de implementação do cluster será de multi-master com etcd (ETCD,) (controlador de logs, e estado do cluster) atachado (KUBERNETES,), ou rodando nos mesmo computadores masters do cluster. Essa arquitetura recomendada, garante a alta disponibilidade do cluster importante para que não haja interrupções inesperadas durante a orquestração das cargas de trabalho (execução, disponibilidade e garantia de estado desejado). Apenas não garantindo uma recuperação rápida de outages dos nós mestres, significando a perda de todos os estados e, com isso, havendo a necessidade de reconfiguração do cluster. Porém considerando os recursos limitados, na justificativa desse projeto, a alta disponibilidade de estado, significaria na obrigatoriedade do mesmo número de computadores disponibilizados para etcd e masters para controle dos estados.

Toda a implementação da solução e os componentes relacionados serão conteinerizados, possibilitando sua orquestração pelo cluster de Kubernetes®. Apenas as configurações dos clusters em si e seu provisionamento não estarão conteinerizados. Esses serão disponibilizados em outro repositório específico. Para ciclo de vida da aplicação e configurações gerais da solução será utilizado uma estratégia de organização de código em monorepo, facilitando a visualização, centralização, sincronização, padronização como benefícios primários e demonstrados na literatura reforçando a adoção dessa estratégia (BRITO; TERRA; VALENTE, 2018).

3.4 Configuração e provisionamento do cluster

Para provisionamento do *cluster* utilizaremos Ansible®da empresa RedHat®. Sua adoção se deu pela característica minimalista de configuração inicial, facilidade de uso e uma característica fundamental que diminui o *overhead* de operação necessária para sua

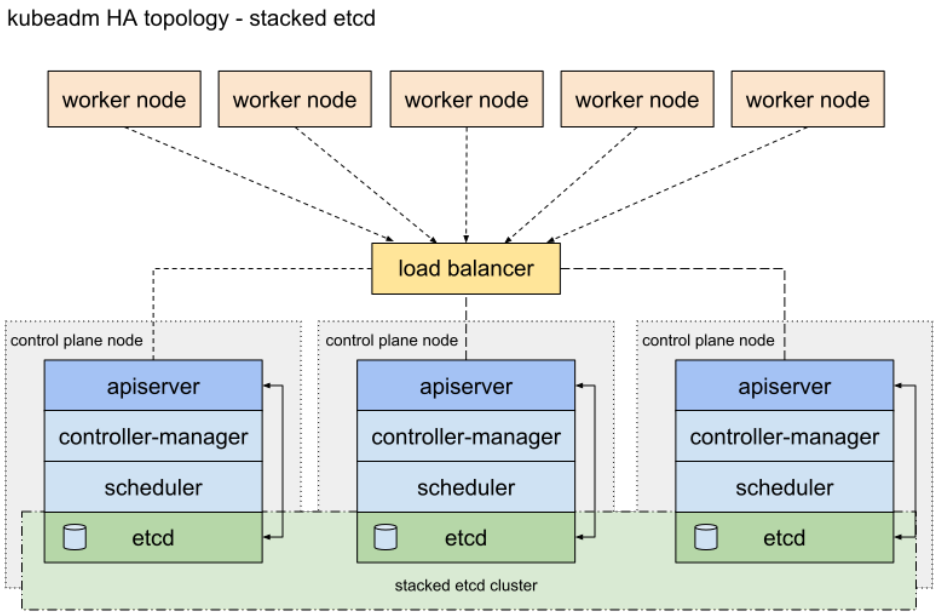

Figura 5 – Kubernetes Arquitetura de alta disponibilidade (Fonte: The Linux Foundation®, 2021)

utilização, não possuir agente instalado nas no inventário de máquinas gerenciadas.

O principal ganho em existente no uso de um sistema de gerenciamento de configuração (CMS) é o não gerenciamento do sistema para utilizá-lo, não há necessidade de configuração ou instalação de nenhum binário específico para sua utilização, reduzindo assim a complexidade de utilização inicial.

A configuração inicial é realizada pela disponibilidade de acesso via rede, python na máquina a ter sua configuração gerenciada (asset, ou recurso) e por meio de alguns tipos de autenticação (kerberos, WinRM, SSH etc), sendo no caso utilizado o protocolo SSH (RFC4254,) por chaves assimétricas o que garante um nível aceitável de segurança, especialmente quando se é possível escolher os algoritmos de criptografia e suas possíveis variações como RSA e ED25519.

3.5 Análise de dados

O banco de dados “Vendas de Medicamentos Controlados e Antimicrobianos - Medicamentos Industrializados”, disponibilizado pelo governo brasileiro (via portal dados.gov.b), será utilizado nesse trabalho. Os anos correspondentes dos dados são no período entre 2014 e 2021 e o banco possui mais de 70 GB e mais de 530 milhões de linhas, sendo assim suficiente para ser utilizado como carga de trabalho ao se calcular uma regressão linear para consumo do medicamento de azitromicina. Como caso base para comparação e teste

do modelo proposto, em termos de tempo para processamento da análise proposta acima, utilizaremos um único computador (desktop) com capacidade equivalente a 8 computadores de menor capacidade (1vCPU e 2GB De RAM) designados com nós do cluster, portanto contendo recursos de 8 vCPUs e 16GB de RAM.

3.6 Monitoramento

Serão utilizados para monitoramento de execução das cargas de trabalho Prometheus®e para visualização dos dados Grafana®, ambos sendo configurados a partir do provisionamento do cluster, ainda com a ferramenta proposta inicialmente Ansible®. Dessa forma será possível avaliar parâmetros de taxa de lotação das máquinas base, pelo parâmetro de processador e memória, operações de leitura e escrita no disco e também tráfego de rede. Por meio dessas ferramentas será possível ainda, avaliar dados de tempo de execução de cargas de trabalho em ambos os ambientes propostos e assim poder compará-los sob eficiência de uso de hardware.

3.7 Comparação entre tipos de virtualização

A arquitetura do tipo x86 foi o tipo mais comum de arquitetura de computadores e boa parte da tecnologia de virtualização inicialmente foi desenvolvida nessa arquitetura (FAYYAD; LucPerneel; TIMMERMAN, 2013). Nesse contexto será também utilizado na construção desse trabalho virtualizações com guest e host OS em x86, para garantir que outros estudos possam ser relacionados na obtenção dos resultados desse trabalho. Como descrito anteriormente, serão utilizadas maquinas virtuais com 2GB de RAM e containers com limitações de mesmo tamanho para a realização de configuração do cluster nesses ambientes, e a partir destes a execução dos workloads.

O tipo de teste de aplicação será um macrobenchmark (system level benchmark) (HUGE, ; SCHEEPERS, 2014) comparando parâmetros de uso de CPU, memória e tempo de execução da carga de trabalho proposta na sessão 3.6 deste trabalho. Os parâmetros serão avaliados tanto na máquina de suporte a virtualização e também nas máquinas virtuais e containers, bem como o tempo provisionamento do container da carga de trabalho, tempo de execução e quantidades de falhas.

Esse método USE (Usage, Saturation and Errors) (GREGG,) será utilizado para extrair e apresentar métricas e avaliar possíveis problemas, sendo que esse não é o foco do trabalho, mas captura de forma adequada os parâmetros de hardware citados acima.

Para controle de tempo será por meio de ferramentas de APM (Application Performance Management) associada a aplicação da carga de trabalho. Dessa forma garantimos a

avaliação de qualidade e possíveis problemas, tempos de execução dentre outros possíveis problemas e métricas de performance da aplicação (TANG et al., 2021).

3.8 Cronograma

A primeira parte do Trabalho de Conclusão de Curso consiste em um estudo dos métodos de virtualização, plataformas elegidas como possíveis soluções e a composição de arquitetura da plataforma de orquestração das cargas de trabalho. Bem como a construção de um modelo inicial de assim como a definição de uma arquitetura de referência a ser refinada no TCC II junto aos demais diagramas do sistema. A solução proposta para realizar tanto a orquestração como também a comparação das cargas de trabalho também serão validadas durante a execução do TCC II. Na Figura 6 é apresentado o cronograma total do projeto:

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	

4 Conclusão

O desenvolvimento do presente trabalho possibilitou a avaliação de diferentes tipos de virtualização para comparação na implementação da plataforma de orquestração de cargas de trabalho de analise de dados em saúde. Através de profunda revisão bibliográfica foram identificados outras plataforma de orquestração de cargas de trabalhos e foi possível fazer uma avaliação critica baseada nos requisitos disponibilizados nas documentações oficiais, bem como a avaliação crítica do proposito ao qual o presente trabalho se propunha. Ainda na literatura, foi possível encontrar dados e informações a respeito dos impactos socio-econômicos resultantes da restrição orçamentária a pesquisa de uma forma geral. Em específico no caso desse trabalho para análise de dados em saúde, especialmente no auxilio de extração de informações relevantes de grande base de dados. O que torna possível a tomada de decisão em saúde pautada em dados e também auxilia a população a ter melhor noção, baseada em dados, da situação de saúde no qual se encontra e assim poder auditar os órgãos públicos responsáveis pela condução do SUS e políticas de saúde associadas.

4.1 Trabalhos Futuros

Na continuação deste trabalho, o provisionamento do ambiente de testes com as devidas restrições e sua consequente avaliação será realizada, coletando dados de tempo de execução, taxa de utilização de memória e CPU, para comparação das duas propostas de sistemas virtualizados. Permitindo assim avaliar qual estratégia mais adequada para simulação de implementação e orquestração de dados. No encerramento desse trabalho possibilita-se que o conhecimento desenvolvido ao longo do TCC possibilite a avaliação de implementação em maquinas físicas, podendo levar ao recrutamento de maquinas heterogêneas e de diversas configurações de hardware afim de criar um ou mais *clusters* com capacidade suficiente para escalonar diversos tipos de cargas de trabalho, a exemplo da desse trabalho afim de possibilitar o processamento de dados massivos.

Referências

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>. Citado na página 4.

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>. Citado na página 1.

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>. Citado na página 9.

CLOUDSTACK. **Installation Guide — Apache CloudStack 4.16.0.0 documentation**. 2022. Disponível em: <https://docs.cloudstack.apache.org/en/latest/installguide/index.html>. Citado na página 5.

COMER, D. **The Cloud Computing Book: The Future of Computing Explained**. [S.l.]: CRC Press, 2021. Google-Books-ID: QCs0EAAAQBAJ. ISBN 978-1-00-038427-7. Citado na página 8.

CORPORATION, I. **Big Data Analytics: Intel's IT Manager Survey on How Organizations Are Using Big Data**. [S.l.], 2012. Citado na página 4.

DIJCKS, J.-P. **Oracle: Big data for the enterprise**. [S.l.], 2013. Citado na página 4.

ETCD. **Install**. Section: docs. Disponível em: <https://etcd.io/docs/v3.5/install/>. Citado na página 9.

FACELI, K. et al. Inteligência artificial: uma abordagem de aprendizado de máquina. 2011. Citado na página 4.

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. Citado na página 11.

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>. Citado na página 1.

GNU. **A Quick Guide to GPLv3 - GNU Project - Free Software Foundation**. Disponível em: <https://www.gnu.org/licenses/quick-guide-gplv3.en.html>. Citado na página 8.

GREGG, B. **The USE Method**. Disponível em: <https://www.brendangregg.com/usemethod.html>. Citado na página 11.

HUGE, M. **Different Types of Benchmarks**. Disponível em: <https://www.cs.umd.edu/users/meesh/cmsc411/website/projects/morebenchmarks/types.html>. Citado na página 11.

KUBERNETES. **Kubernetes Documentation | Kubernetes**. Disponível em: <https://kubernetes.io/docs/home/>. Citado na página 9.

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. Citado na página 4.

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>. Citado na página 1.

MEHTA, N.; PANDIT, A. Concurrence of big data analytics and healthcare: A systematic review. **International Journal of Medical Informatics**, 2018. Citado na página 4.

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>. Citado na página 5.

OPENSTACK. **OpenStack Docs: Xena Installation Guides**. 2022. Disponível em: <https://docs.openstack.org/xena/install/>. Citado na página 5.

PLATFORM definition and meaning | Collins English Dictionary. 2022. Disponível em: <https://www.collinsdictionary.com/dictionary/english/platform>. Citado na página 5.

RESENDE, L.; VIANA, L.; VIDGAL, P. **PROTOCOLOS CLÍNICOS DOS EXAMES LABORATORIAIS**. 1. ed. Minas Gerais: Secretaria de Estado de Saúde de Minas Gerais, 2009. Citado na página 4.

RFC4254. Disponível em: <https://datatracker.ietf.org/doc/html/rfc4254>. Citado na página 10.

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. Citado na página 11.

SCHROECK, M. et al. Analytics: el uso de big data en el mundo real. **IBM Institute for Business Value, Oxford, Informe ejecutivo**, 2012. Citado na página 4.

TANG, Y. et al. A systematical study on application performance management libraries for apps. **IEEE Transactions on Software Engineering**, IEEE, 2021. Citado na página 12.

TRUYEN, E. et al. A Comprehensive Feature Comparison Study of Open-Source Container Orchestration Frameworks. **arXiv:2002.02806 [cs]**, mar. 2021. ArXiv: 2002.02806. Disponível em: <http://arxiv.org/abs/2002.02806>. Citado 2 vezes nas páginas 5 e 6.

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. Citado na página 6.