UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

ANÁLISE DE DADOS UTILIZAND *CLUSTER* E BAIXO CUSTO

Tendências de consumo da azitromicina no Brasil antes e durante a pandemia da COVID-19

Felipe Fonseca Rocha

Orientador: Ítalo Fernando Scotá Cunha

Universidade Federal de Minas Gerais

09 de Fevereiro de 2022

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

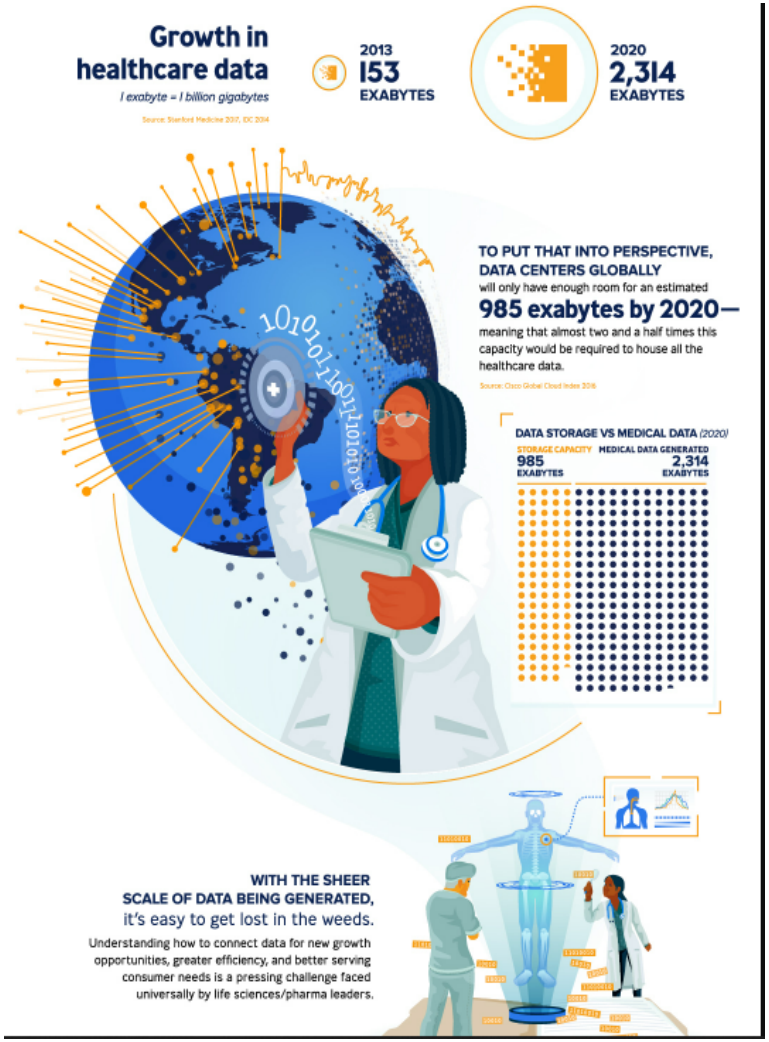
Sumário

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

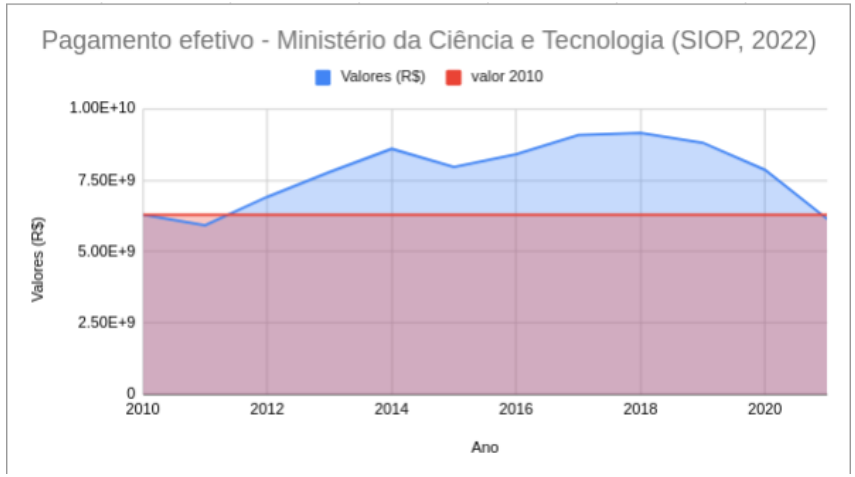
Uso de ferramentas e recursos já disponíveis de maneira correta

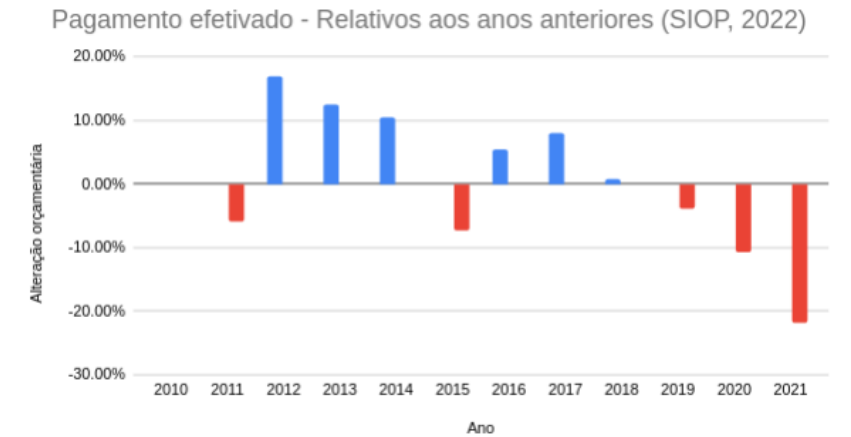
Justificativa Social

- 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 publicas

Justificativa Econômica

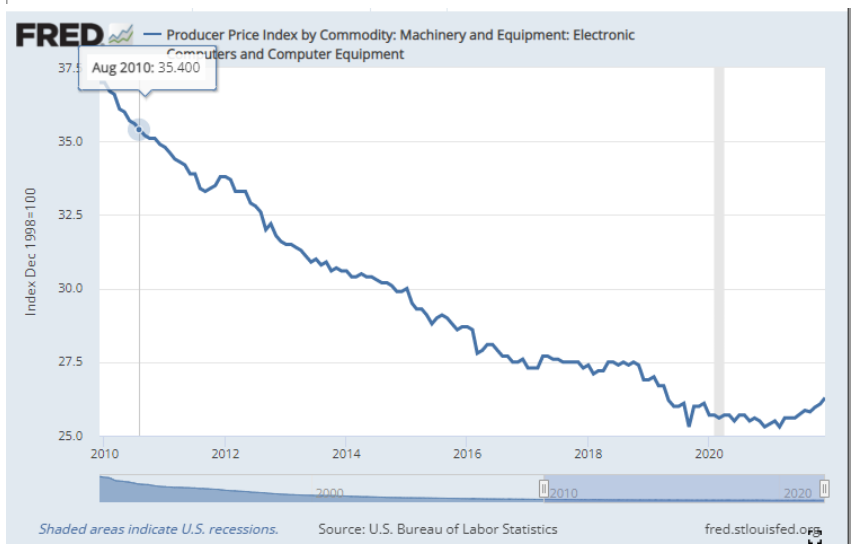
- Gasto na disponibilização dos dados
- Diminuição de verbas para ciência e tecnologia $-2,32\%$





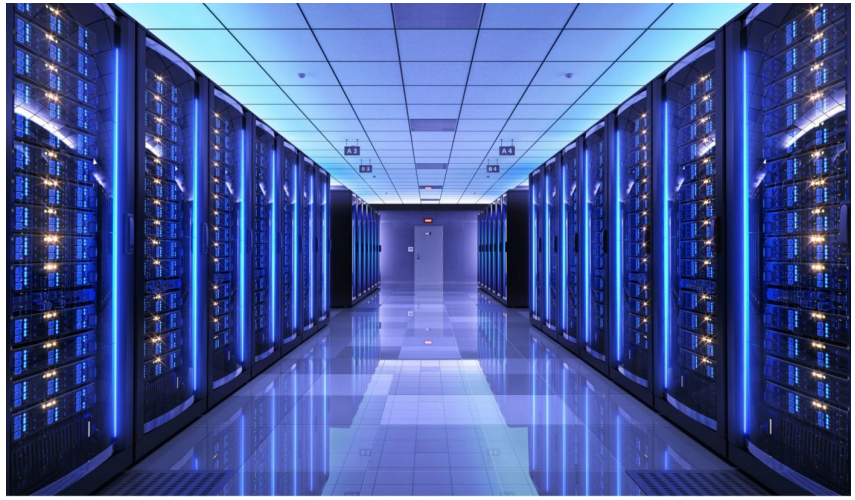
Justificativa Econômica

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



Justificativa Técnica

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



Objetivo I

Objetivos Geral:

Avaliar a viabilidade de orquestração de recursos em *cluster* de baixo custo em ambientes containerizados, para o processamento e a análise dos dados.

Objetivos Específicos:

- Realizar a orquestração de recursos em *cluster* de baixo custo;
- Avaliar tempo de provisionamento, tempo de execução e disponibilidade do cluster;
- 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;
- Disponibilzar um cluster pronto para uso para UFMG, bem como ferramentas de auxilio no provisionamento;

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

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 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

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

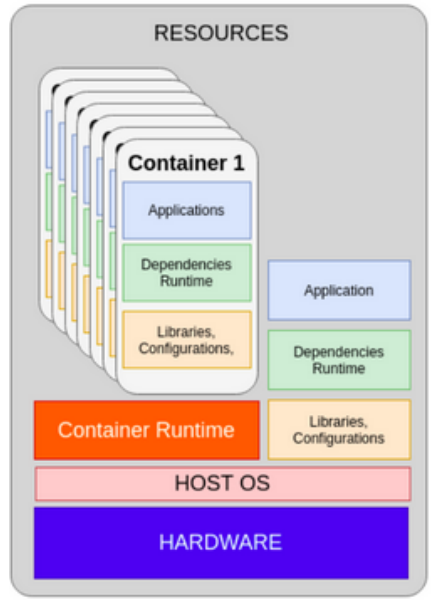
Alternativas *open source*

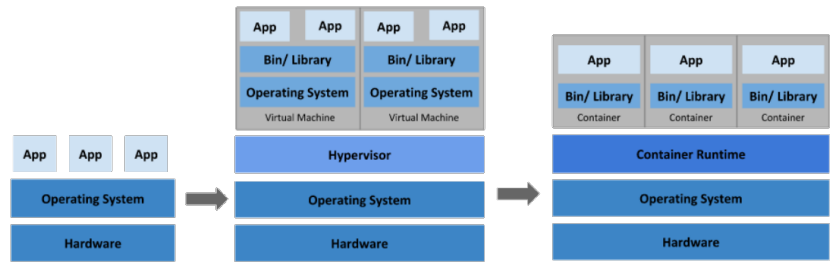
- Considerando
  - O escopo deste trabalho
  - As estratégias para processamento e análise de dados disponíveis no mercado

  As soluções encontradas no mercado foram agrupadas em dois grupos:
  - Soluções de Computação em nuvem privada:
    - Se estendem para além do proposito 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®

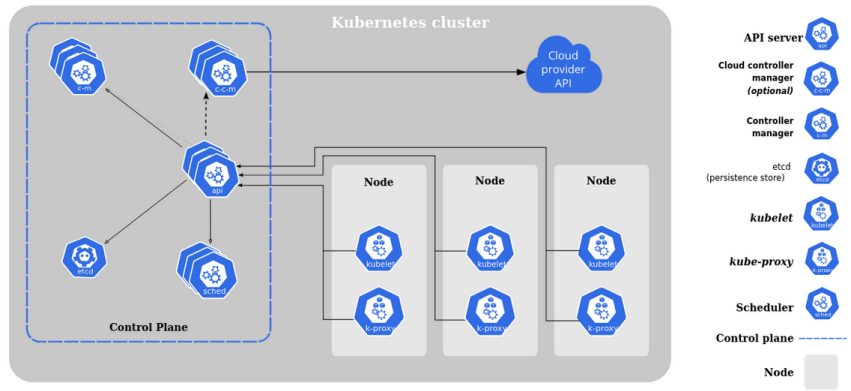




UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

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



Abordagem I

Utilizar u *Cluster* Kubernetes® como plataforma de orquestração de cargas de trabalho em computadores desktops.

- Cargas de trabalho:
  - Analise de tendencia de uso de azitromicina entre 2014 e 2021
- Composição do cluster com computadores *desktops* reaproveitados
- Minimizar trabalho local e priorizar a possibilidade de provisionamneto remoto
- redução do CAPEX e otimizar utilização de hardware ocioso ou subutilizado
- reaproveitamento de maquinas

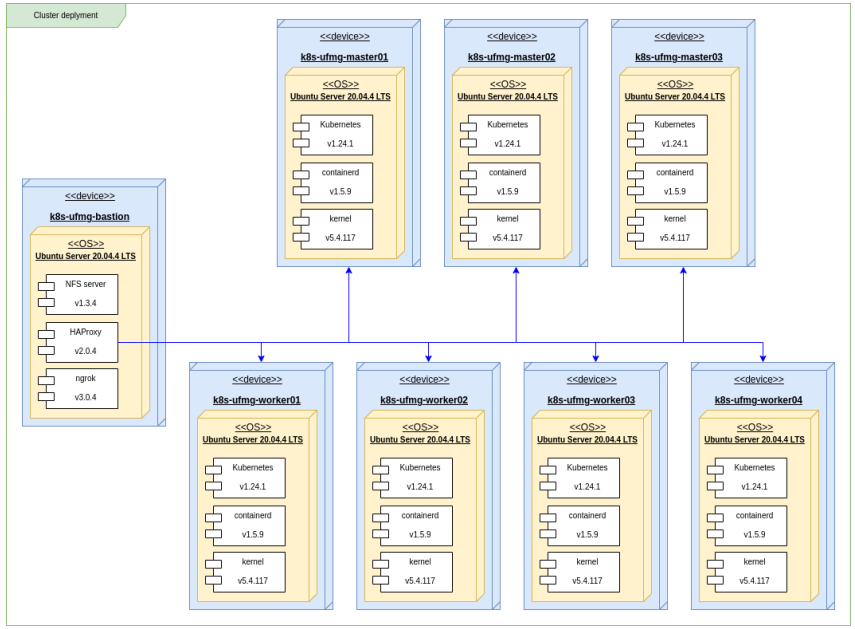
Abordagem

O uso de conceitos e metodologias de DevOps:

- CI (integração contínua)
- CD (entrega contínua)
- Monitoramento
  - método USE, parâmetros de utilização, saturação e erro
  - avaliação de utilização dos nós durante processamento

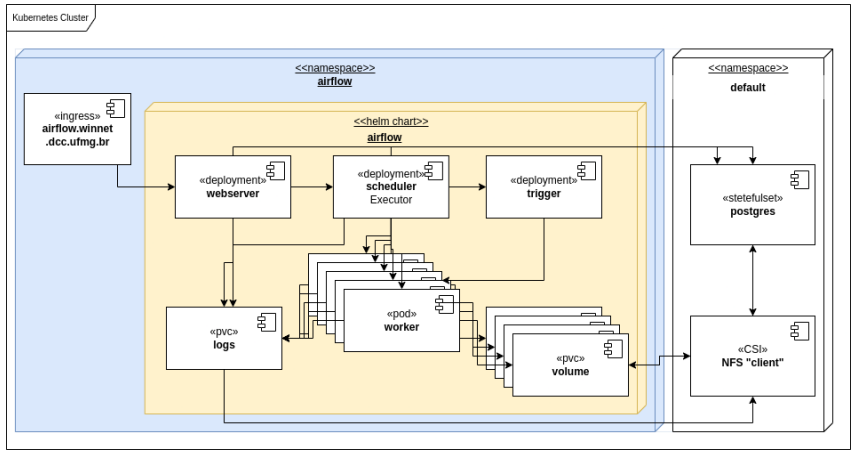
Especificações I

- Cluster:
  - Composição:
    - 4 computadores com 6 CPUs e 8GB de RAM (*load balancer e control-plane*)
    - 4 computadores com 6 CPUs e 16GB de RAM (*workers*)
  - Containers para processamento e análise:
    - arquitetura: **amd64**
    - 1 vCPU
    - 2 GB de RAM
    - 90 containers (1/mês de análise) [procesamento]
    - 1 container / usuário [análise]



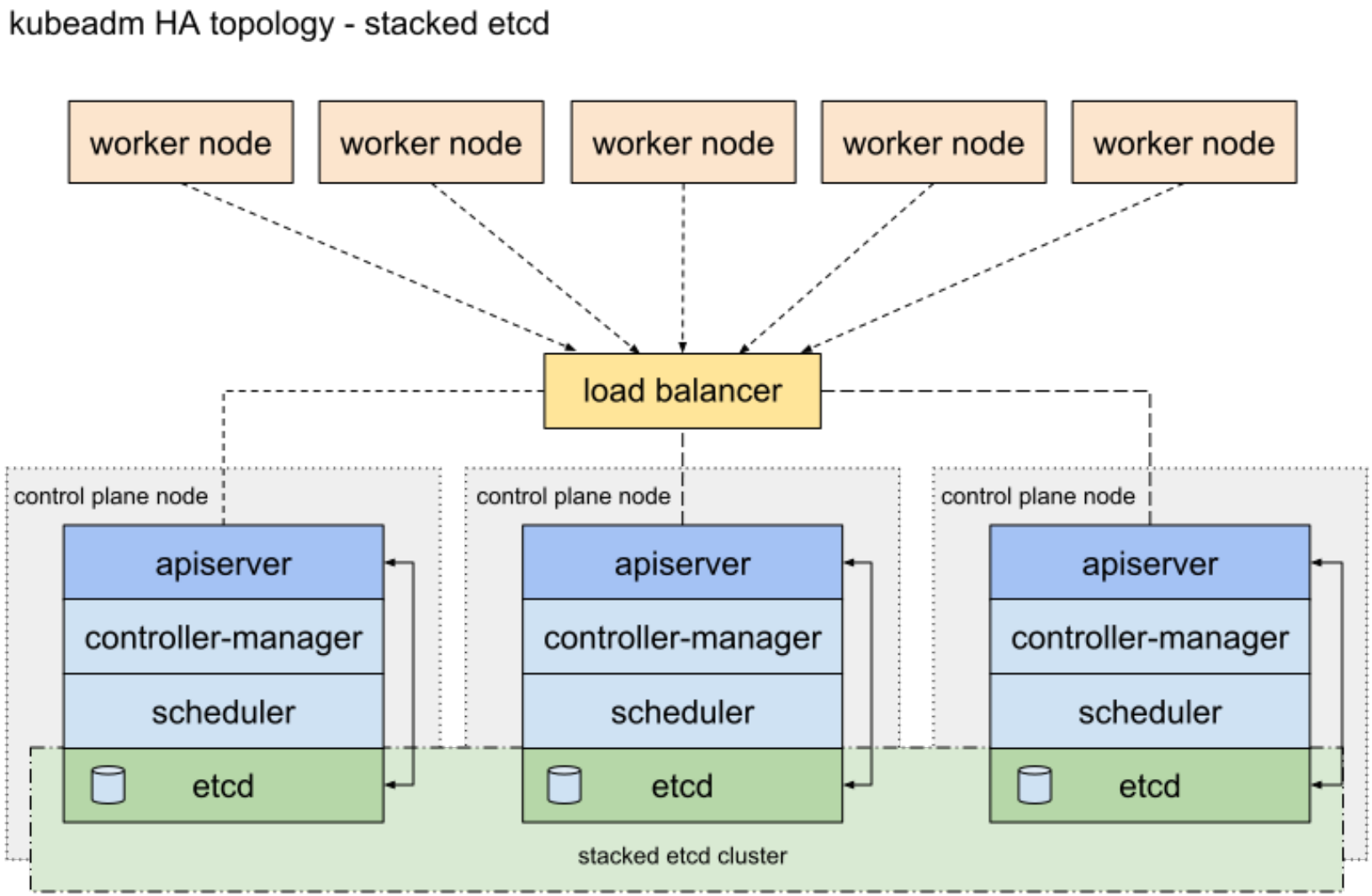
Especificações II

- Orquestração do processamento dos dados originias:
  - Apache Airflow®
  - Kubernetes executor
  - Python Operators
- Consumo e análise de dados tratados:
  - JupyterHub - gerenciamento de notebooks
  - Jupyter Notebooks - análise dos dados

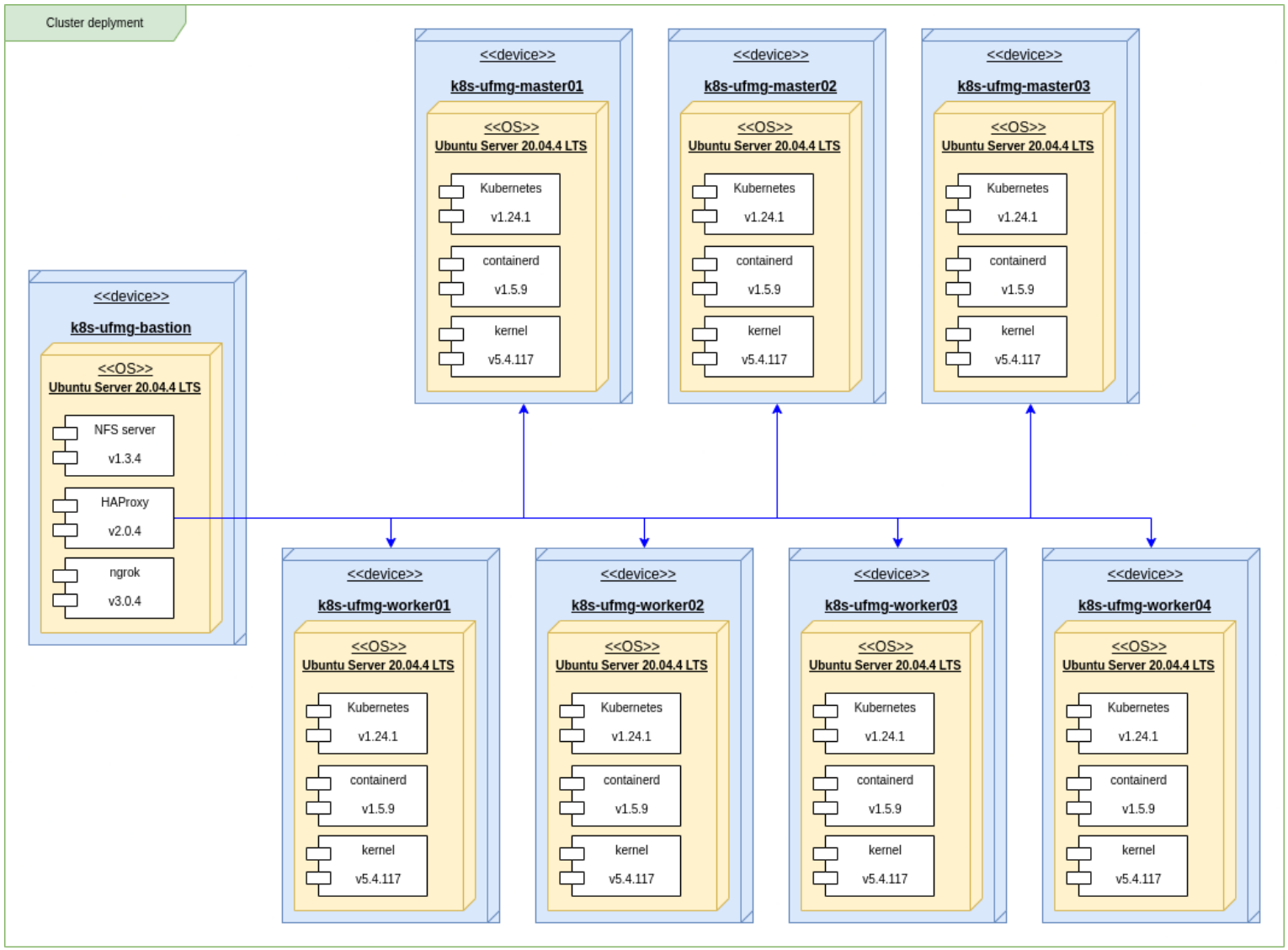


UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

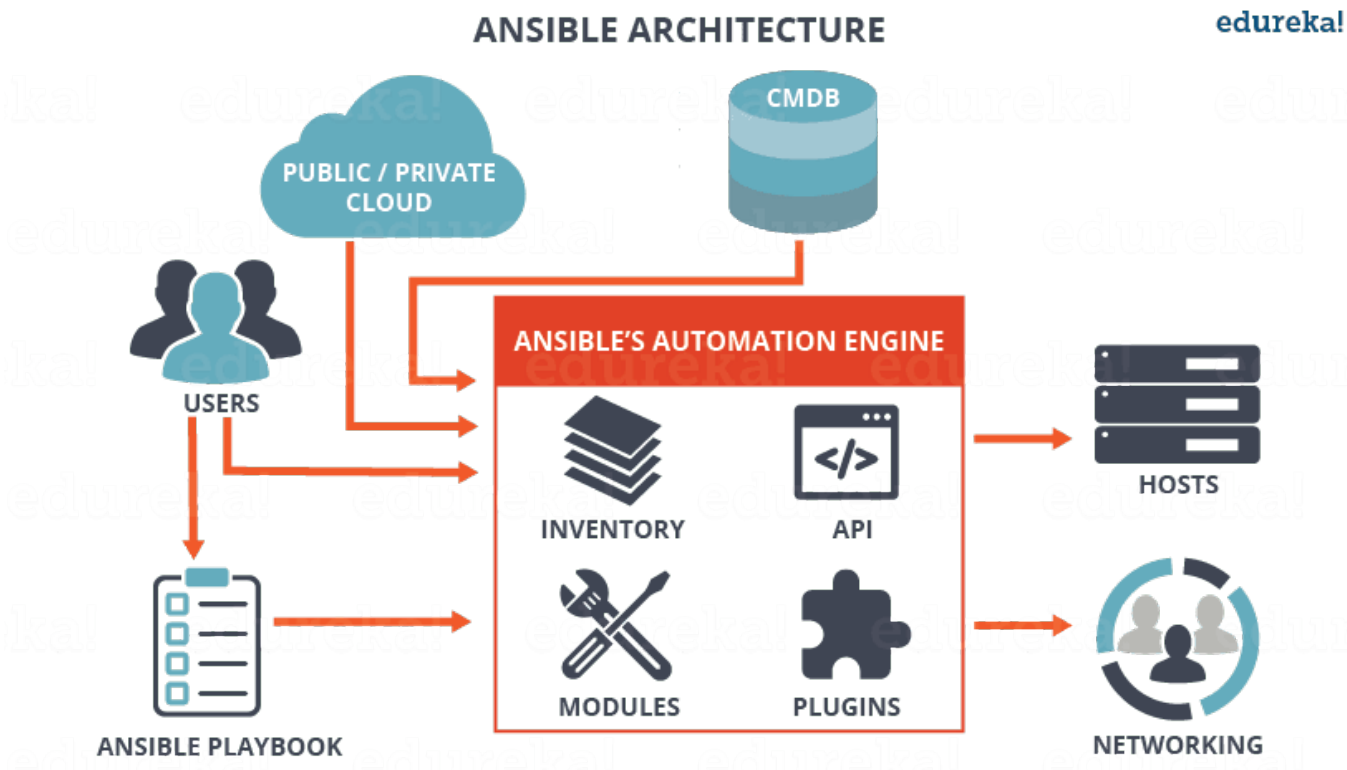
Arquitetura Orquestrador



Arquitetura Orquestrador

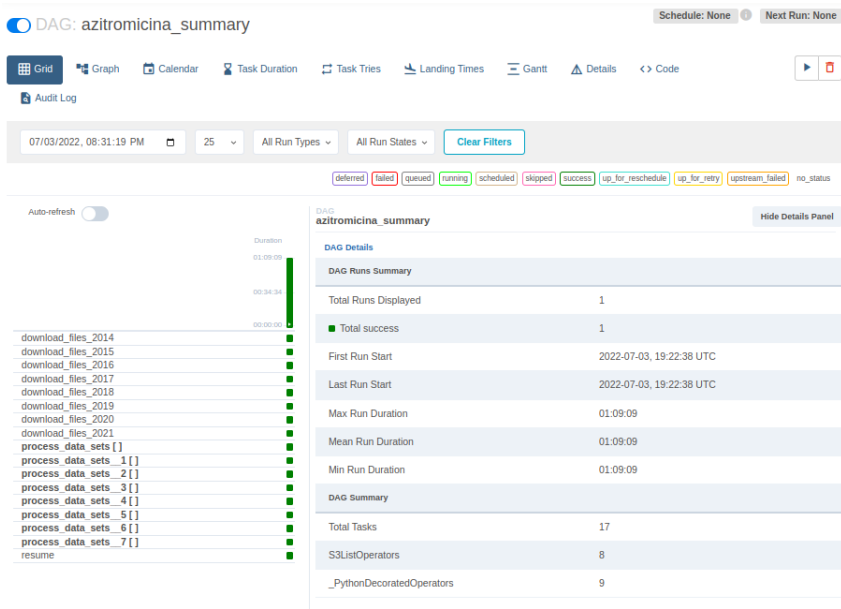


Gerenciamento de configuração



Monitoramento

- *Node Exporter* Expor métricas de Host
- *Prometheus* - Monitoramento de sistemas e Banco de dados de series temporais
- *Grafana* - Dashboard e observ-abilidade
- Airflow - Relatório de tempo de execução, falhas, tentativas



UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Avaliação de utilização do cluster I

- macrobenchmark (system level benchmark) - Teste utizando 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

  Tempo de Implementação:
- Tempo de configuração do cluster

  Método base utilizado para coleta de informações:
- Metodo USE de avaliação (Checklist Linux)

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 por região
- Análise de tendência do consumo de azitromicina no país
- Avaliação compartiva de 2 anos anteriores ao COVID-19

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

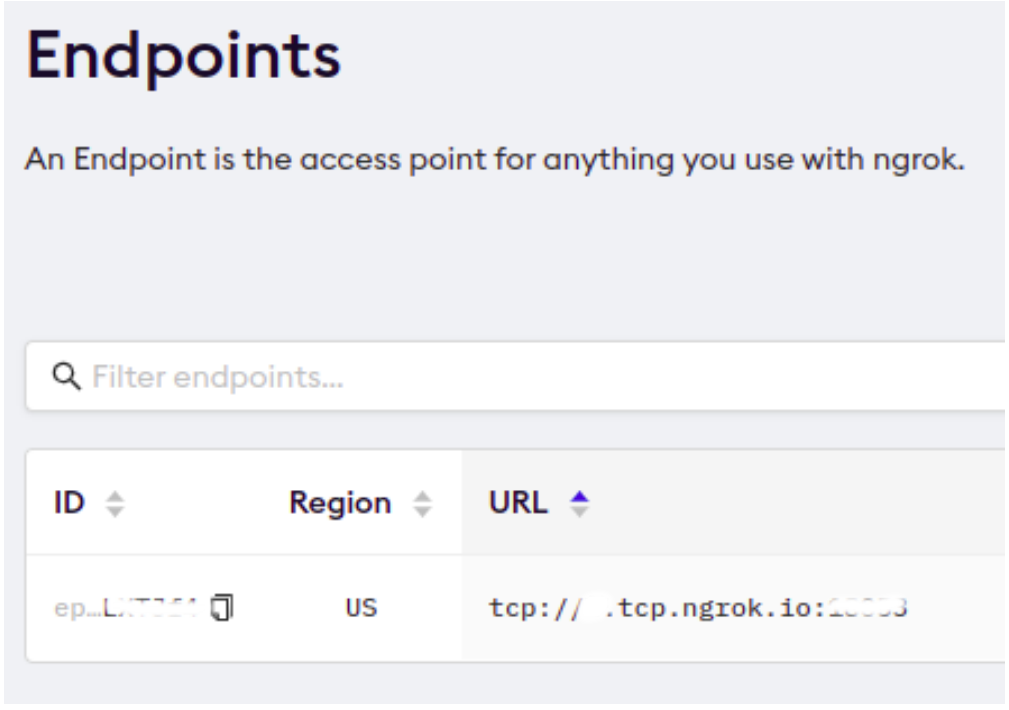
Ondas				Dates			Cascata
Fases de Projeto	Atividades do TCC	Atividades	Objetivos	S	Data inicial	Data Final	
Exploratório	Proposta TCC I	Elaboração de estratégias de busca	Identificar estudos parecidos, explorar tecnologias disponíveis e avaliar oportunidades e conceitos associados aos usuários	1	17/10/2021	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.	24	06/03/2022	02/04/2022	
		Instrumentação (software) e verificação		26	03/04/2022	16/04/2022	
		Implementação da análise e verificação		28	17/04/2022	30/04/2022	
		Testes de Validação		34	01/05/2022	11/06/2022	
Utilização & Suporte		Coleta dos resultados	Captação da utilização em cenário real em projeto de pesquisa parceiro	35	12/06/2022	18/06/2022	
		Discussao dos resultados obtidos.		36	19/06/2022	25/06/2022	
Encerramento		Definição de proximas etápas	Estudo do caso de uso e sumarização dos resultados para apresentação da solução junto a banca	38	26/06/2022	09/07/2022	
		Formalização dos trabalho e apresentação		39	10/07/2022	16/07/2022	

Resultados e discussões

- Provisionamento
  - Tempo de configuração inicial
    - sem imagem personalizada: 2 dias
    - cloud-init: 2h (possivel redução se utilizado imagens em rede)
  - Tempo de configuração cluster
    - configuração total manual: ¿¿ 12h
    - ansible: 15 min
    - helm (deploy aplicação) + terraform (orquestração de deploy): 10-20 min
- Execussão dos job:
  - Tempo serial com limite de 2GB de RAM e 1CPU 90 pods: 1h 30m
  - Tempo com orquestrador em paralelo com limite de 2GB de RAM e 1CPU 90 pods: 53m
- Lead Time:
  - Tempo de deploy Airflow DAG: 3-6 min
  - Tempo de deploy JupyterLab: 4-10 min
-

Disponibilidade dos recursos deste trabalho

Todos os componentes definidos neste trabalho estarão contidos em um repositório público Github, sob a licença pública geral GNU versão 3, para livre acesso.


Endpoints
An Endpoint is the access point for anything you use with ngrok.
Filter endpoints...
ID
Region
URL
US
tcp://
.tcp.ngrok.io:

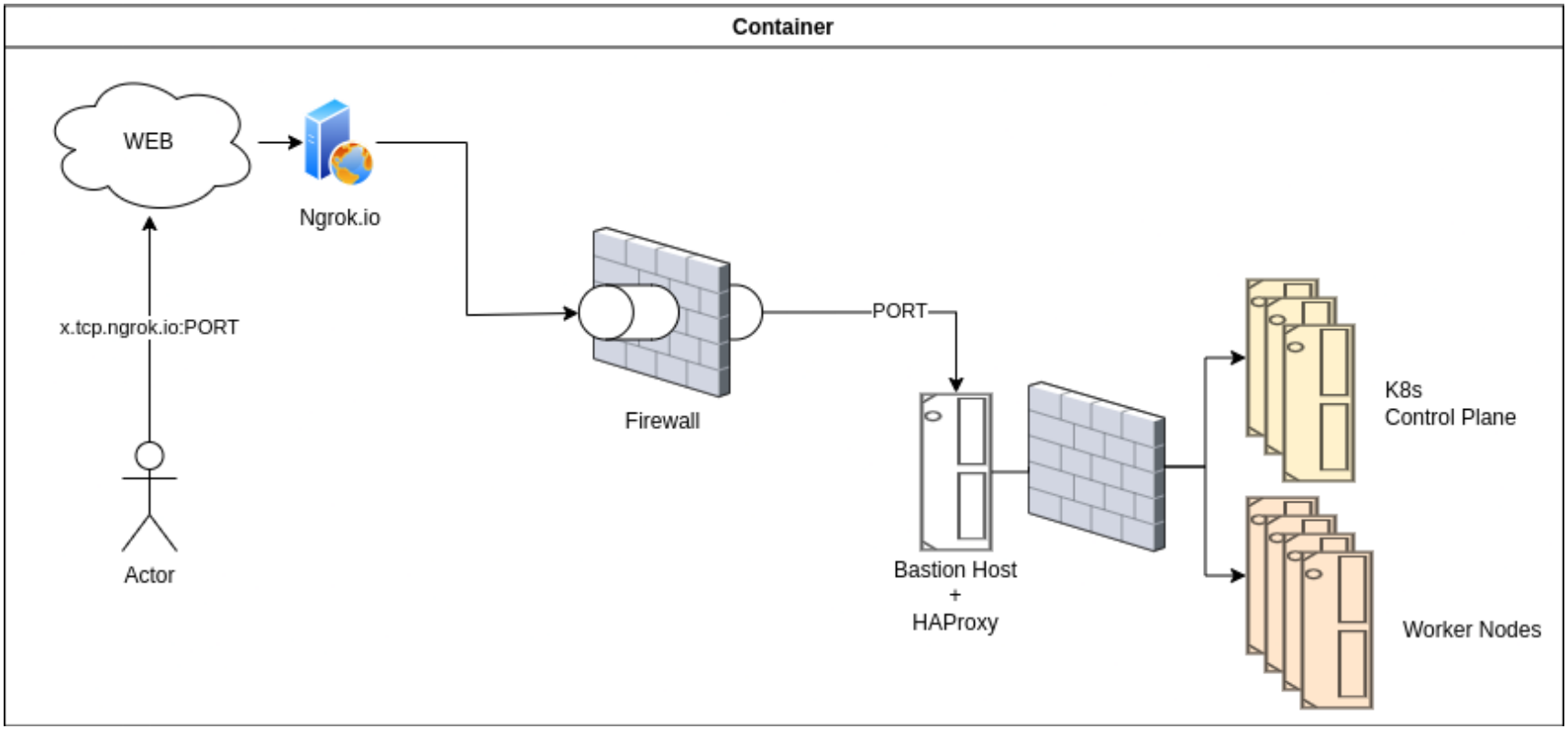


Figura: Funcionamento NgRok

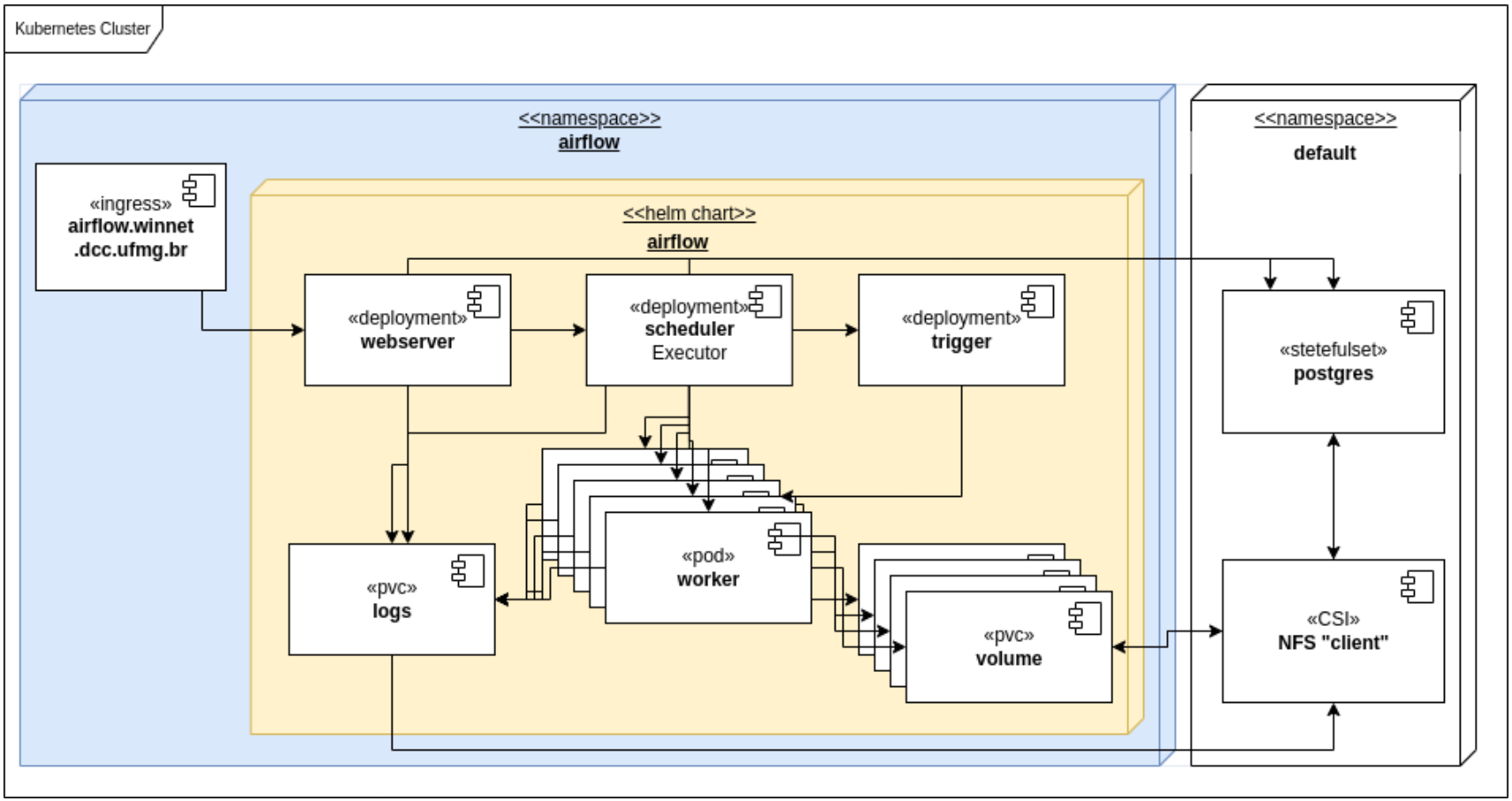


Figura: Ariflow - Diagrama de Deploy

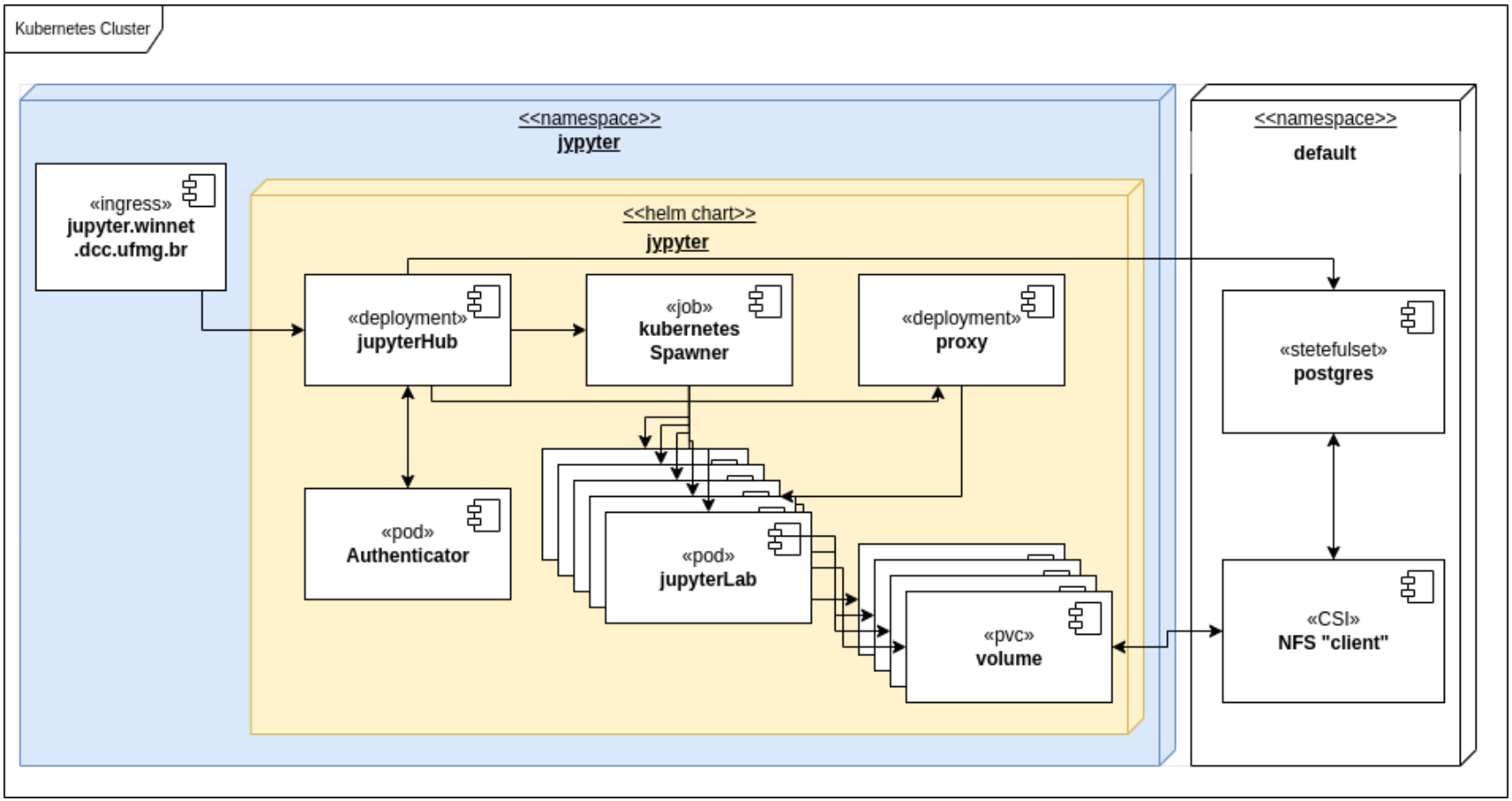


Figura: Jupyter - Diagrama de Deploy

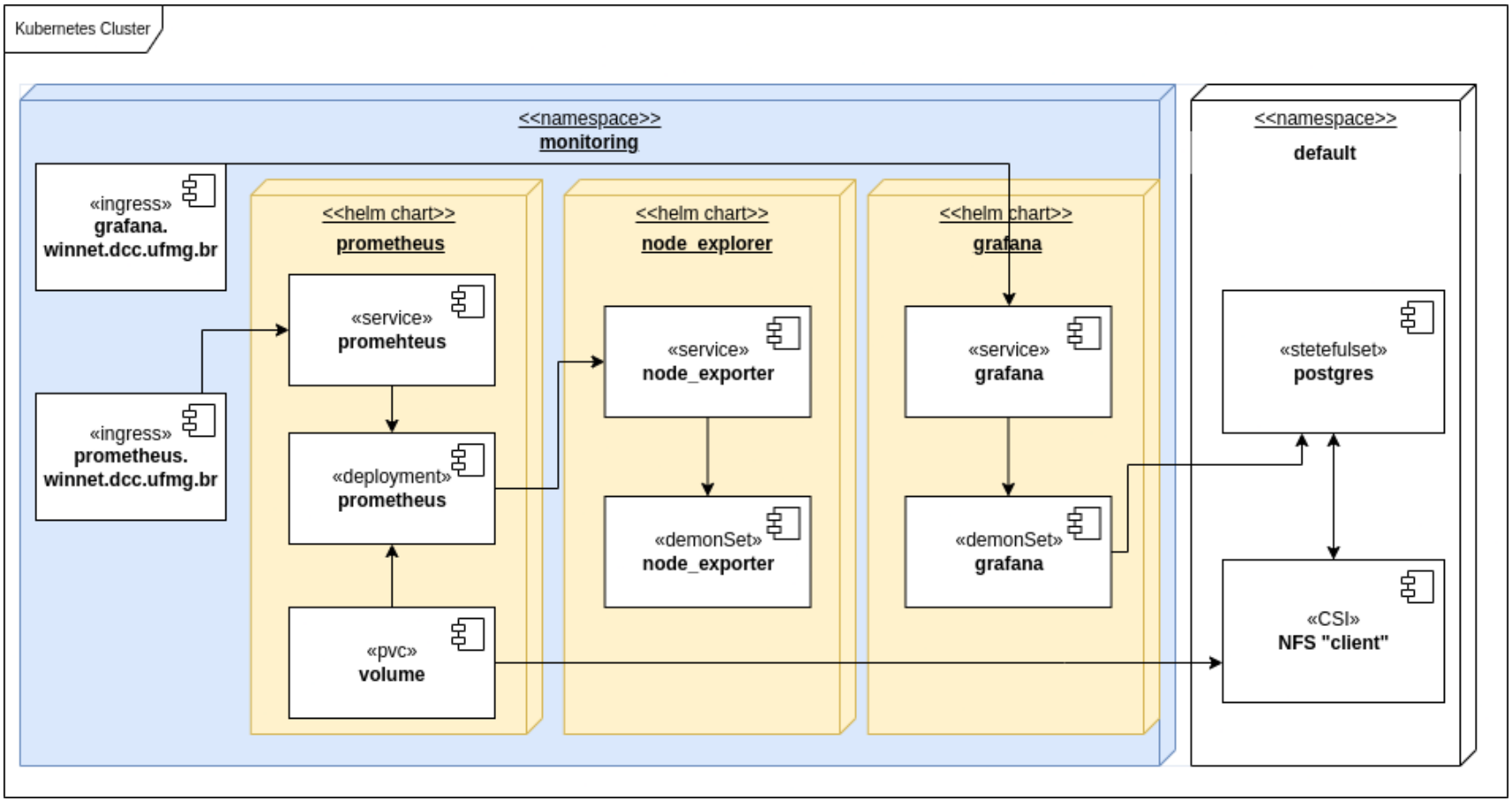


Figura: Monitoramento - Diagrama de Deploy

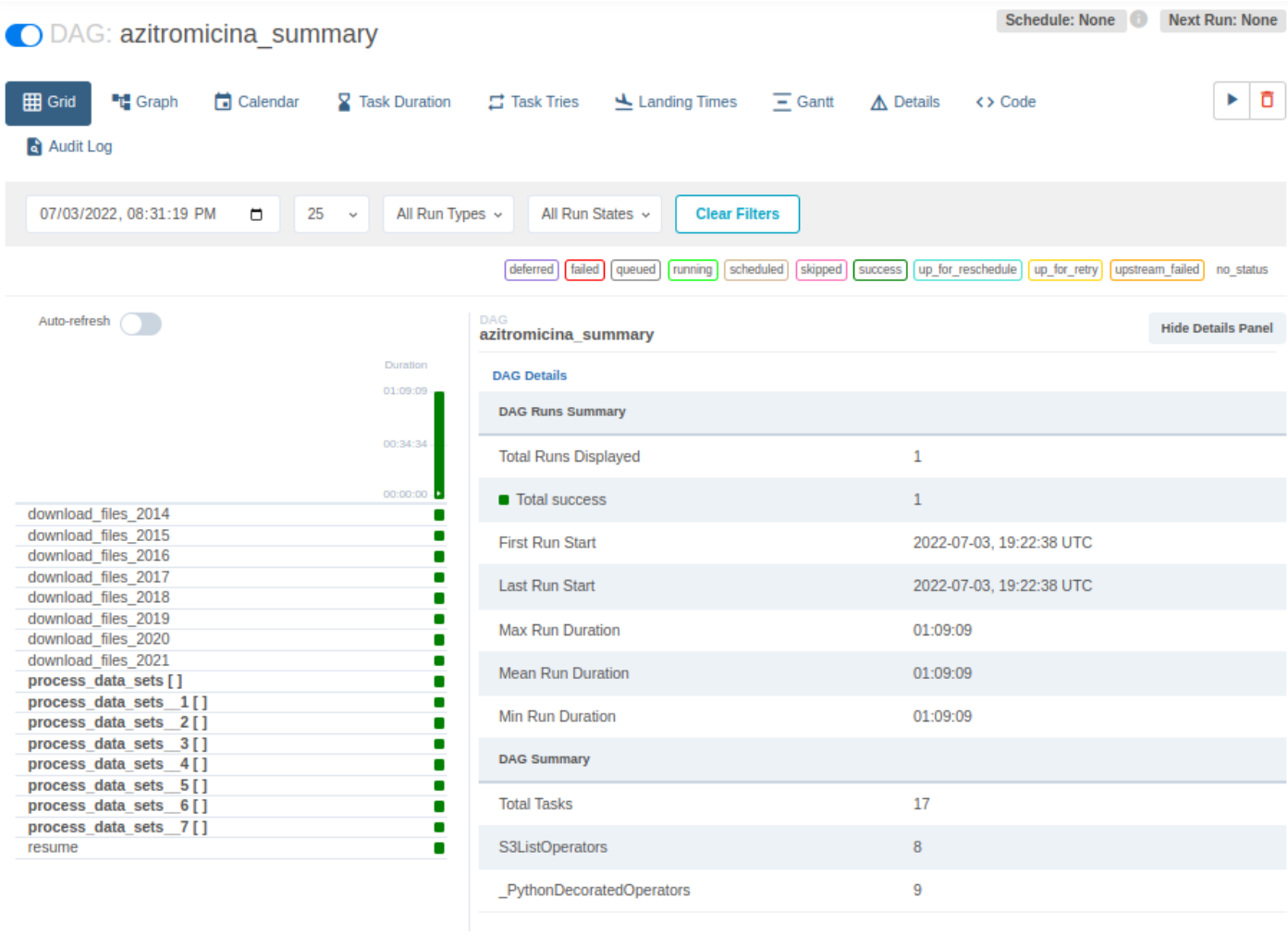


Figura: Relatório de orquestração

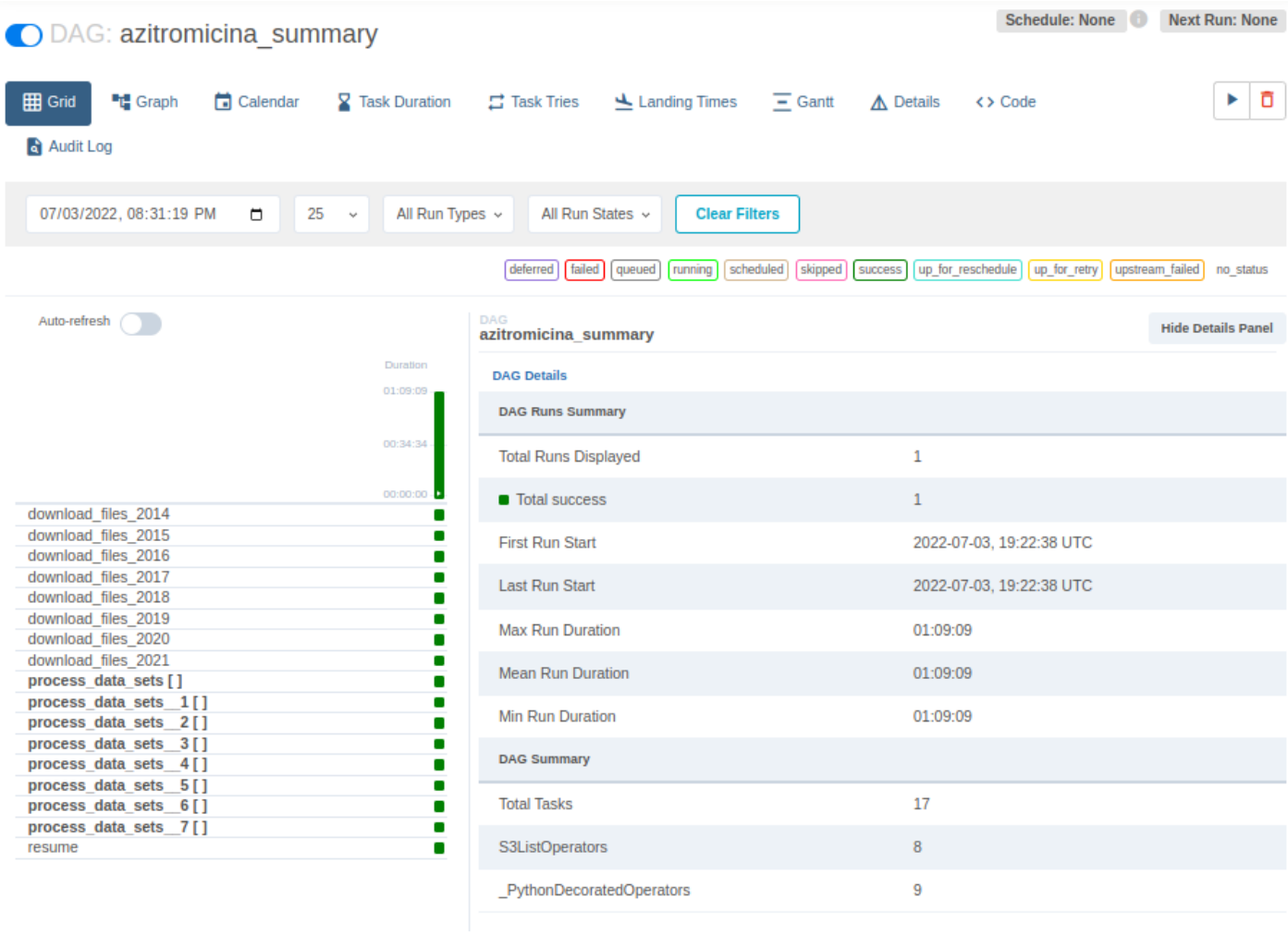


Figura: Relatório de Orquestração

Figura: Monitoramento execução

Figura: Monitoramento execução 2

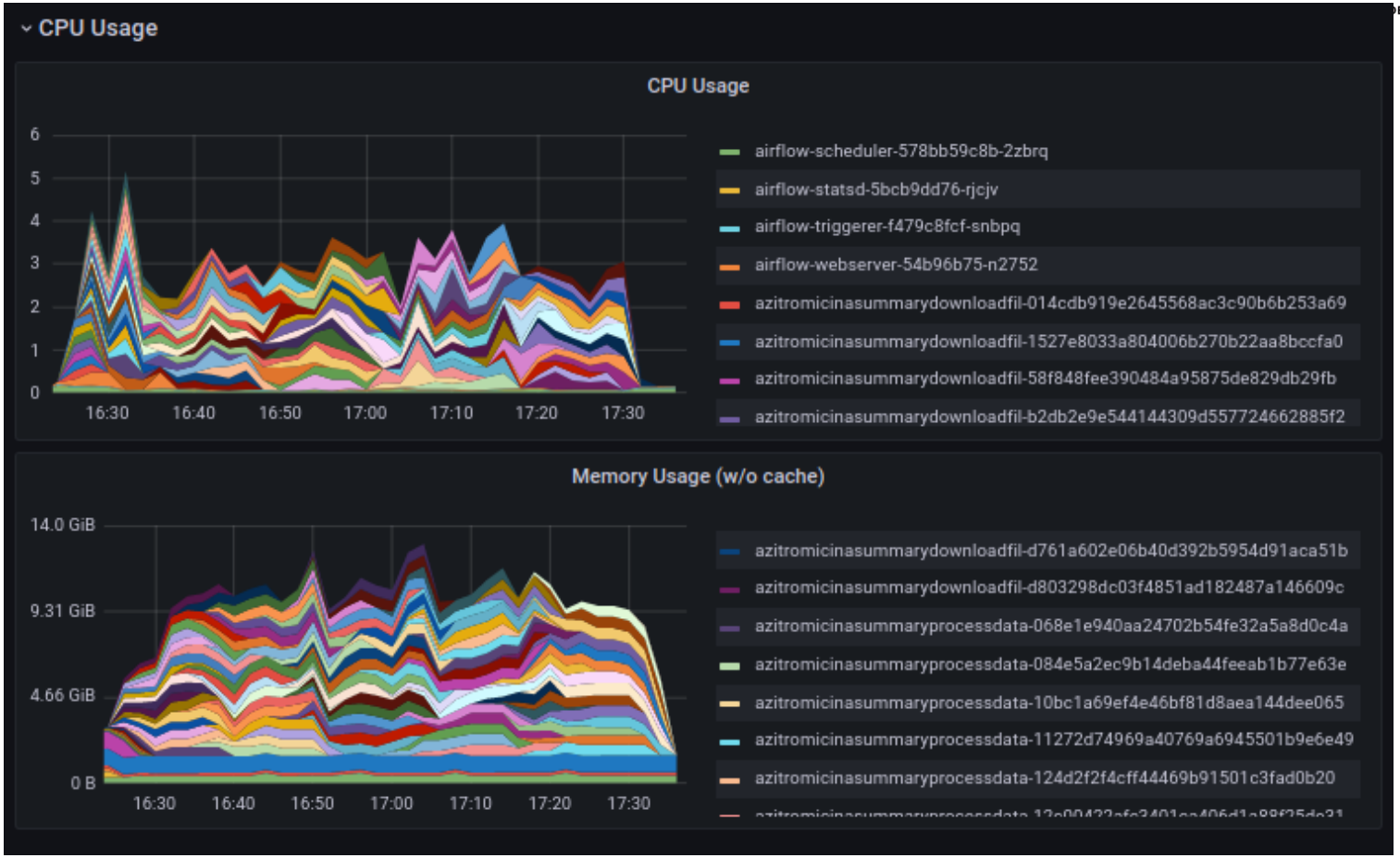


Figura: Monitoramento execução 3

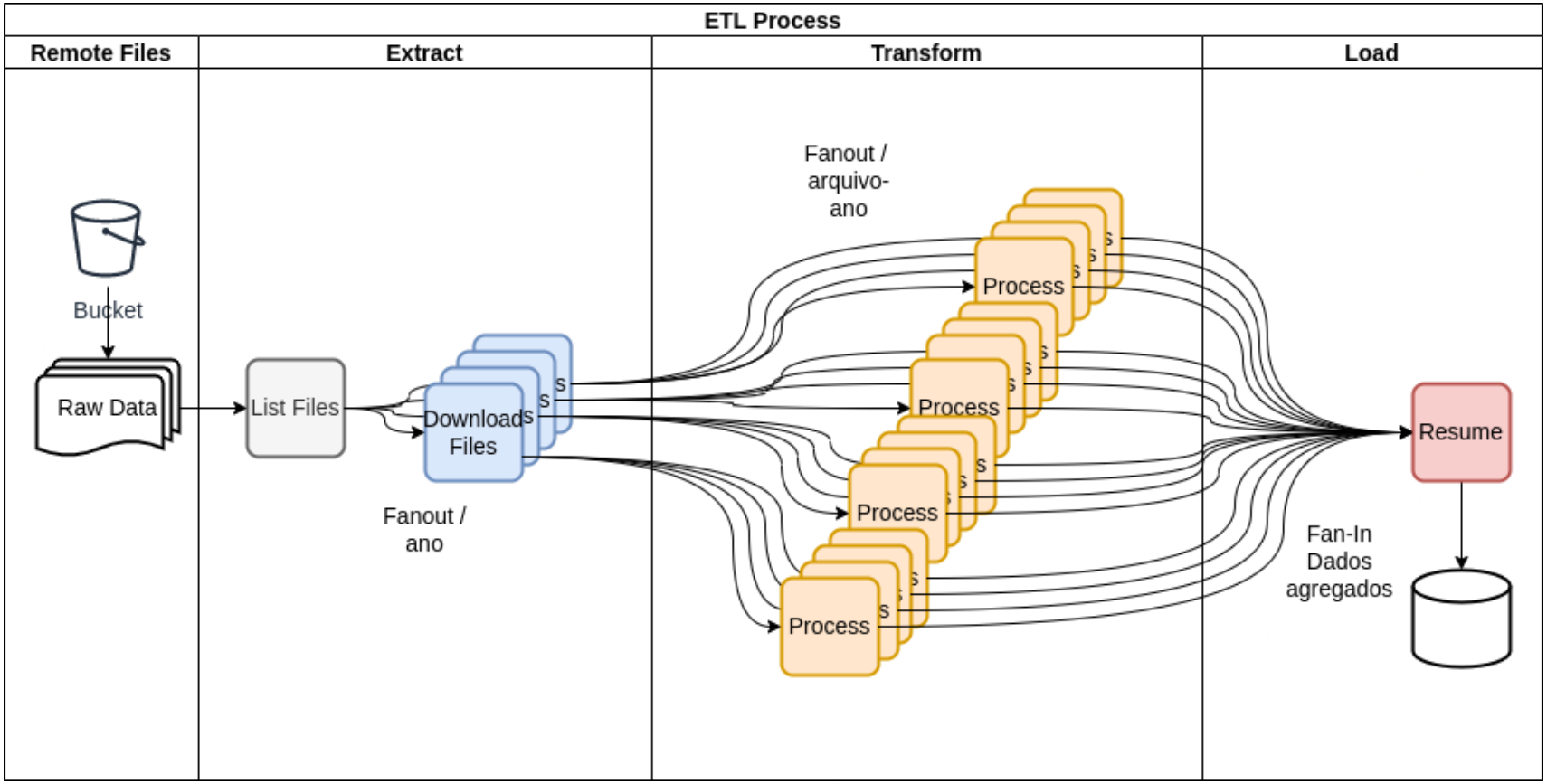


Figura: Ansible inventory

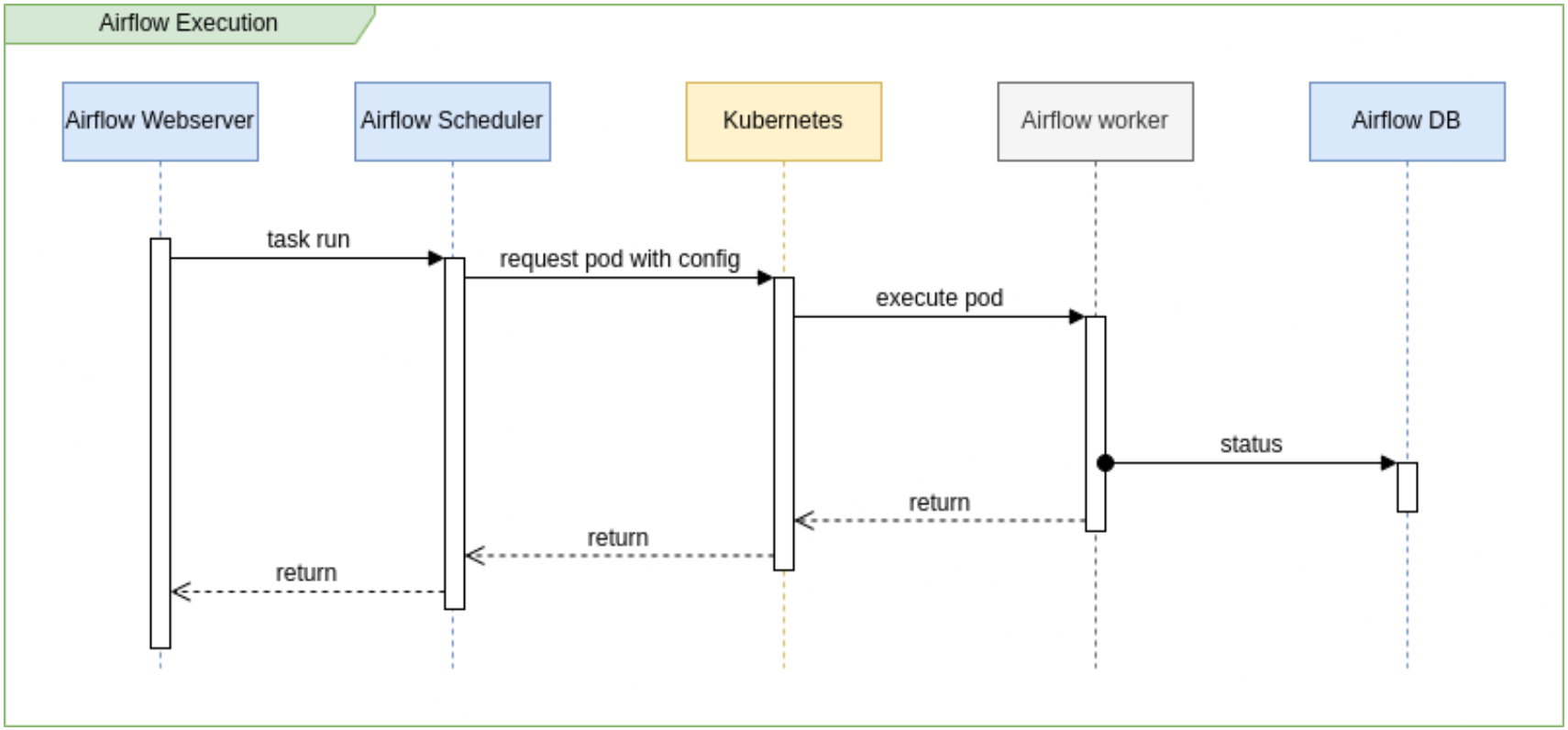


Figura: Airflow - Diagrama de Sequencia

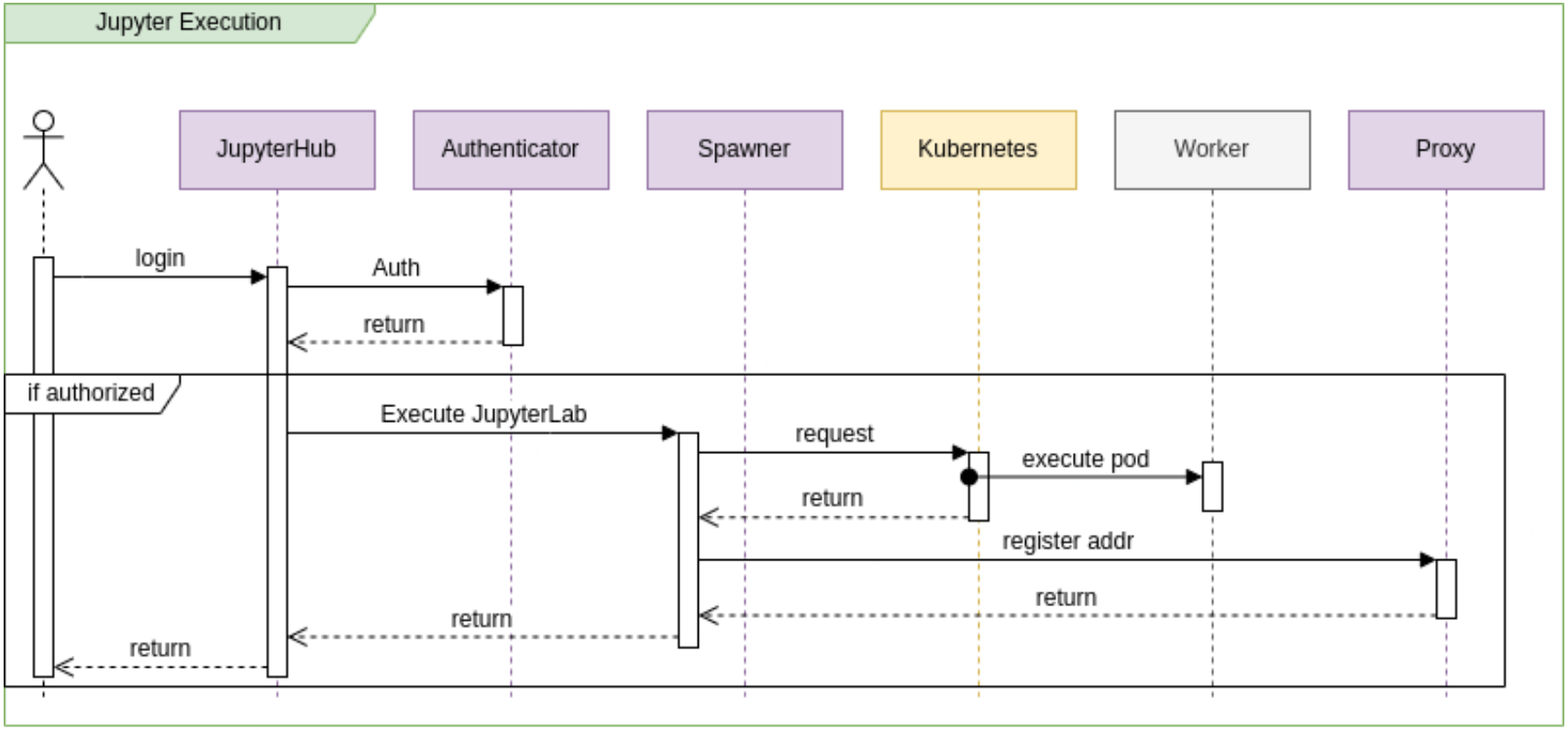


Figura: Jupyter - Diagrama de Sequencia

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

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 maquina base.

Referências I

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

OBRIGADO
:)