

Atividade 5 – Tolerância a Falhas e Monitoramento com Kubernetes + Prometheus

Objetivo Geral

Implantar uma aplicação distribuída com Kubernetes em um ambiente local, aplicando técnicas de **tolerância a falhas**, **auto-recuperação**, **escalonamento horizontal** e **monitoramento remoto**, utilizando dois notebooks interligados.

Descrição da Implantação

✓ Notebook A – Cluster Kubernetes

- Criar um cluster local com **Minikube**.
- Implantar uma aplicação com múltiplos pods, iniciando com **duas réplicas**.
 - *Atenção: a aplicação não pode ser reutilizada de atividades anteriores da disciplina.*
- Habilitar o mecanismo de **auto-healing**, garantindo que pods sejam recriados automaticamente em caso de falha.
- Configurar o **Horizontal Pod Autoscaler (HPA)** com base no uso de CPU.

✓ Notebook B – Prometheus

- Implantar o **Prometheus** para monitoramento remoto do cluster Kubernetes no Notebook A.
 - Exibir métricas em tempo real, como:
 - Número de pods ativos
 - Uso de CPU
 - Estado dos pods (**Running**, **Failed**, **Pending**)
 - Ações disparadas pelo HPA
-

Demonstração de Tolerância a Falhas

Antes de iniciar os testes, mostre que sua aplicação está rodando normalmente com as duas réplicas previstas.

1. Deleção Manual de Pod (Auto-Healing)

- Delete manualmente um dos pods da aplicação.
- Observe como o controlador do Kubernetes detecta a falha e recria automaticamente um novo pod.

📌 O que demonstrar:

- A recriação rápida de um novo pod após a deleção.

- O status de “Terminating” do pod anterior e a entrada do novo em “Running”.
- A atualização das métricas no Prometheus (mudança no número de pods ativos, novo identificador, tempo de reação).

2. Sobrecarga de CPU (Escalonamento Horizontal)

- Gere uma carga de CPU artificial em um ou mais pods da aplicação.
- Observe como o HPA aumenta automaticamente o número de réplicas para atender à demanda.

O que demonstrar:

- A elevação do consumo de CPU no Prometheus.
- A criação de novos pods, respeitando o limite configurado no HPA.
- O tempo de resposta entre o pico de CPU e o escalonamento automático.
- A redução do número de réplicas após estabilização (se aplicável).

Entrega no SIGAA

Você deverá entregar um **relatório em PDF** com as seguintes características:

- Texto entre **4 e 6 páginas**, escrito de forma **clara, objetiva e direta**.
- Capturas de tela com **legendas explicativas**.
- Evidências diretas: use **prints de comandos, métricas e dashboards**, tanto do terminal quanto do Prometheus.