

**Espaço CMaker**

**Fazendo Fazedores e Criando Conexões**

RELATÓRIO PARCIAL

**Modularização e clusterização de IA**

**CURITIBA**

**23/07/2025**

#### Vinicius Baldan Herrera

**Modularização e clusterização de IA.**

|  |  |
| --- | --- |
|  |  |

**CURITIBA  
10/01/2024**

**SUMÁRIO**

[1 INTRODUÇÃO 4](#_Toc57830827)

[2 OBJETIVO(S) 4](#_Toc57830828)

[3 MATERIAIS E MÉTODOS 4](#_Toc57830829)

[4 RESULTADOS PARCIAIS 5](#_Toc57830830)

[4.1 PROCESSOS E/OU PROCEDIMENTOS DE DESENVOLVIMENTO TECNOLÓGICO 5](#_Toc57830831)

[5 ETAPAS FUTURAS 5](#_Toc57830832)

[6 REFERÊNCIAS 5](#_Toc57830833)

[ANEXOS 6](#_Toc57830834)

# 1 INTRODUÇÃO

O crescimento do uso de Inteligência Artificial (IA) tem exigido infraestruturas computacionais cada vez mais robustas. Em muitos ambientes de inovação, como espaços makers e laboratórios educacionais, os recursos disponíveis são limitados e muitas vezes compostos por máquinas heterogêneas e de gerações distintas. Isso gera desafios técnicos para obter desempenho satisfatório, além de dificultar a escalabilidade quando novas demandas surgem.

A modularização e a clusterização de sistemas de IA surgem como estratégias para otimizar o uso dos recursos existentes, permitindo que máquinas diferentes trabalhem de forma coordenada, escalando o poder de processamento conforme a necessidade. Por meio de tecnologias como contêineres (Docker) e orquestradores de tarefas, é possível criar uma camada de abstração que facilita a integração e manutenção do ambiente de IA, mesmo em infraestruturas modestas.

Este trabalho propõe o desenvolvimento de um cluster de IA modular, composto por computadores de diferentes capacidades, com o objetivo de avaliar desempenho, escalabilidade e eficiência operacional. A proposta busca demonstrar que, mesmo com hardware heterogêneo, é possível alcançar um ambiente funcional para execução de modelos de IA, explorando práticas modernas de virtualização e containerização.

**2 OBJETIVO(S)**

Projetar e implementar uma infraestrutura modular para execução de modelos de IA, baseada em clusterização de máquinas heterogêneas utilizando contêineres, visando escalabilidade e melhor aproveitamento de recursos disponíveis.

Podem-se destacar os seguintes objetivos específicos:

i. Levantar e caracterizar os recursos computacionais disponíveis.

ii. Implementar contêineres para padronizar ambientes de execução de IA.

iii. Avaliar métricas de desempenho (latência, throughput) e escalabilidade em cenários de carga variável.

iv. Documentar a arquitetura desenvolvida e propor melhorias futuras.

v. Avaliar a aceitação e a eficácia do agente no cotidiano do Espaço Maker.

# 3 MATERIAIS E MÉTODOS

\*\*Sugestão\*\*

Materiais utilizados

Para o desenvolvimento do projeto de conteinerização e modularização de IA, foram utilizados computadores e servidores disponíveis no espaço de pesquisa, sendo previamente inventariados para compreender a capacidade de CPU, memória RAM e GPUs disponíveis. Esses equipamentos formaram a infraestrutura base para execução e testes.

Foi utilizado software de conteinerização, como Docker e Podman, para criação de ambientes isolados que garantem portabilidade e consistência entre diferentes máquinas. Para gerenciar a execução distribuída, foi empregado um orquestrador de contêineres, a exemplo do Docker Swarm ou do Kubernetes, permitindo escalar tarefas e distribuir cargas de trabalho de forma eficiente entre nós do cluster.

Foram preparados e utilizados modelos de IA de teste, aplicados em inferências de imagens, textos e pequenos modelos de linguagem (LLMs). Além desses, foram incluídos modelos de IA previamente selecionados para execução local, como redes neurais dedicadas a classificação ou geração, permitindo comparar diferentes tipos de carga e comportamento de execução.

Métodos aplicados

O primeiro passo foi o mapeamento dos recursos disponíveis. Nesta etapa, foi realizado um inventário detalhado de cada máquina envolvida no projeto, especificando processadores, quantidade de núcleos, memória RAM disponível, presença ou ausência de GPU e estimativa de consumo energético. Isso permitiu identificar limitações físicas e preparar um plano adequado para distribuição das cargas de IA.

Na sequência, ocorreu a padronização com contêineres. Foram criadas imagens Docker contendo ambientes de IA pré-configurados, incluindo dependências, frameworks e bibliotecas específicas, garantindo que todos os nós do cluster pudessem executar as mesmas tarefas sem problemas de compatibilidade. Essa padronização também facilitou o versionamento e a atualização de modelos.

Após a padronização, foi realizada a implementação do cluster. Nesta etapa, o orquestrador de tarefas escolhido (Swarm ou Kubernetes) foi configurado para distribuir automaticamente a execução dos modelos de IA entre os nós disponíveis. Foram criadas regras de balanceamento e mecanismos de monitoramento para acompanhar o status de cada nó e otimizar o uso dos recursos.

Com o cluster funcional, foram conduzidos testes de carga e escalabilidade. Os modelos de IA foram executados sob diferentes níveis de demanda, medindo-se tempo de resposta, utilização de CPU/GPU e throughput. Essa análise permitiu avaliar a robustez do sistema e identificar gargalos potenciais.

Por fim, foi feita uma análise aprofundada dos resultados. Os dados coletados durante os testes foram estudados para identificar pontos de melhoria, tanto na modularização quanto no balanceamento de carga. Essa análise forneceu insumos para ajustes na arquitetura e nas configurações do orquestrador, buscando melhor desempenho geral do sistema. \*\*Sugestão\*\*

# 4 RESULTADOS PARCIAIS

Considerando o cronograma proposto no plano de trabalho, apresentar o desenvolvimento da pesquisa e os resultados atingidos até o momento. Se ocorreu alguma dificuldade ou alteração de cronograma da pesquisa, deve ser aqui relatado e justificado.

## 4.1 PROCESSOS E/OU PROCEDIMENTOS DE DESENVOLVIMENTO TECNOLÓGICO

Projetos de PIBITI têm como objetivo adicional, práticas conectadas ao desenvolvimento teórico aplicado conectados ao desenvolvimento de tecnologia e inovação. Dessa forma, descrever o processo de produção do artefato (produto ou processo) até o momento. Em que momento ele se encontra ou se há alguma geração de alternativas para sua produção.

# 5 ETAPAS FUTURAS

Considerando o cronograma proposto no plano de trabalho, analise a adequação das próximas etapas ao tempo disponível para conclusão.

Não se esqueça de submeter seu texto a uma criteriosa revisão gramatical e ortográfica. Bom trabalho!

# 6 REFERÊNCIAS

Apresente as referências **CONFORME GUIA PARA NORMALIZAÇÃO DE TRABALHOS ACADÊMICOS ABNT.**

# ANEXOS

**OUTRAS ATIVIDADES REALIZADAS**