

## Estrutura do Projeto

```
/
├── api_chat_bd/ # API para upload e busca de PDFs
├── chat_api_llm/ # API para integração com Gemini
├── chat_interface/ # Interface web Streamlit
├── chat_bd/ # Configuração Qdrant
├── charts/myapp/ # Helm Chart
├── scripts/ # Script de build automatizado
└── README.md
```

## 1. Componentes da Aplicação

### Chat Interface (Streamlit)

- Função: Interface web para interação com usuário
- Recursos: Upload de PDFs, chat com IA, busca de documentos
- Porta: 8501
- Imagem: chatapp/chat-interface:latest

### API Chat BD (FastAPI)

- Função: Upload e processamento de PDFs, busca vetorial
- Recursos: Integração com Qdrant, criação de coleções
- Porta: 8000
- Imagem: chatapp/api-chat-bd:latest

### API Chat LLM (FastAPI)

- Função: Integração com Gemini AI
- Recursos: Geração de embeddings, processamento de perguntas
- Porta: 8000
- Imagem: chatapp/chat-api-llm:latest

### Qdrant (Vector Database)

- Função: Banco de dados vetorial para busca semântica
- Portas: 6333 (HTTP), 6334 (gRPC)
- Imagem: qdrant/qdrant:latest

## 2. Artefatos Kubernetes Utilizados

### Namespace

- chatapp - Isolamento lógico da aplicação

### Deployments

- api-chat-bd
- chat-api-llm
- chat-interface
- qdrant

### Services

- api-chat-bd (ClusterIP:8000)
- chat-api-llm (ClusterIP:8000)
- chat-interface (ClusterIP:8501)

- qdrant (ClusterIP:6333)

#### Ingress

- Classe: nginx
- Host: k8s.local
- Rotas:
  - / → chat-interface:8501
  - /api/bd → api-chat-bd:8000
  - /api/llm → chat-api-llm:8000

#### PersistentVolumeClaims

- api-chat-bd-uploads-pvc - Armazenamento de uploads (1Gi)
- qdrant-pvc - Armazenamento do banco vetorial (1Gi)

#### ServiceAccount

- chatapp-sa - Conta de serviço para os pods

#### ConfigMaps/Secrets

- Variáveis de ambiente configuradas via Helm values

## 3. Pré-requisitos

- Docker (versão 20.10+)
- Minikube (versão 1.30+)
- Kubectl (versão 1.25+)
- Helm (versão 3.10+)

# Instalação e Deploy

## 1. Configurar Minikube

# Iniciar Minikube

```
minikube start --driver=docker
```

# Habilitar Ingress

```
minikube addons enable ingress
```

## 2. Build e Deploy Automatizado

# Dar permissão de execução

```
chmod +x scripts/build_and_load.sh
```

# Executar build e deploy

```
./scripts/build_and_load.sh
```

## 3. Verificar Deploy

# Verificar pods

```
kubectl get pods -n chatapp
```

# Verificar services

```
kubectl get svc -n chatapp
```

# Verificar ingress

```
kubectl get ingress -n chatapp
```

## Acesso à Aplicação

1. <http://k8s.local> # Interface principal
2. <http://k8s.local/api/bd> # API Chat BD
3. <http://k8s.local/api/llm> # API Chat LLM