Kubernetes Resource Planning for Chatbot Application

Overview

This document outlines the Kubernetes cluster resource requirements for deploying a chatbot application composed of multiple microservices, storage systems, and retrieval-augmented generation (RAG) components. The plan is designed for a production-grade, enterprise-level deployment and excludes backup or disaster recovery (DR) considerations.

1. Application Components

Storage Systems

- Postgres
- MinIO
- Qdrant

Monitoring Stack

- Prometheus: Metrics collection
- Grafana: Dashboard visualization
- Jaeger: Distributed tracing
- Alertmanager: Alert handling and routing

Queueing System

- Kafka and Zookeeper: Distributed event streaming platform

Microservices

Core Services:
- adminservice
- authservice
- chatservice
- workerservice
- socketservice
- kong service (API Gateway)

RAG Services:
- cleaning service
- embedding service
- document processing service
- llm model service

2. Resource Estimates

Stateful Components

Component	Resources per Pod	Notes
Postgres	4 vCPU / 8 GB RAM	Consider managed DB
MinIO	4 vCPU / 8 GB RAM	I/O heavy
Qdrant	4 vCPU / 8 GB RAM	Vector DB
Kafka + ZK	4 vCPU / 12 GB RAM	Dedicated node recommended

Core Microservices (including Monitoring Stack)

Service	Resources	Notes
adminservice	1 vCPU / 1 GB RAM	UI/backend low usage
authservice	2 vCPU / 2 GB RAM	Handles login/auth
chatservice	2 vCPU / 3 GB RAM	Main logic engine
workerservice	4 vCPU / 4 GB RAM	Async tasks
socketservice	2 vCPU / 2 GB RAM	WebSocket handling
kong service	2 vCPU / 2 GB RAM	API gateway + rate limiting
prometheus	2 vCPU / 2 GB RAM	Metrics scraper and storage
grafana	1 vCPU / 1 GB RAM	Visualization frontend
jaeger	2 vCPU / 2 GB RAM	Tracing collection and visualization
alertmanager	1 vCPU / 1 GB RAM	Notification and alerting handler

RAG Services

Service	Resources	Notes
cleaning service	1 vCPU / 1 GB RAM	Preprocessing texts
embedding service	8 vCPU / 16 GB RAM	GPU for better speed
document processing	4 vCPU / 8 GB RAM	PDF/OCR/Doc parsing
LLM model service	8–16 vCPU / 32–64 GB RAM	External or GPU preferred

3. VMs Allocation for Enterprise-Grade Production

Node Purpose	vCPU	RAM	Disk	Description
Postgres Node	4	8 GB	512 GB SSD	Dedicated DB node
Qdrant Node	4	8 GB	1 TB SSD	Vector DB workloads
MinIO Node	4	8 GB	1 TB SSD	High I/O throughput
Kafka Node	4	12 GB	100 GB SSD	Kafka broker + zookeeper
Microservices Node	16	64 GB	100 GB SSD	Runs all app and RAG microservices
LLM Model Node	16	64 GB	100 GB SSD	GPU is required to run LLM models
Control Plane Node	4	16 GB	100 GB SSD	Kubernetes master/control plane

Total of (52 vCPU, 180RAM, ~3TB SSD)
Notes:

- We may require additional resources based on customer usage and future application requirements.
- Number of vim should be an even number as discussed earlier so if this is the case we will require an additional vm and separate the RAG micro services described above in an independent vm