

Vamshi Krishna Chatla

vamshikrishnachatla@gmail.com | +91-8639378225 | LinkedIn | LeetCode | Chennai, India

Summary

Software Engineer with 3+ years of experience designing and developing scalable, high-performance Java backend systems and AI-powered services at Zoho. Skilled in Spring Boot, microservices, multithreading, Kafka, Redis, and MySQL, with strong foundations in distributed systems, system design, and data structures. Specialized hands-on experience in LLM orchestration, Retrieval-Augmented Generation (RAG), and search relevance engineering for enterprise SaaS products.

Technical Skills

- **Languages:** Java, Python, SQL
- **Java Backend:** Spring Boot, Spring MVC, Spring Data JPA, Hibernate, REST APIs, Microservices, Multithreading
- **Data, Caching & Messaging:** MySQL, Redis (Cluster, Consistent Hashing), Apache Kafka
- **Cloud & DevOps:** AWS, Docker, Kubernetes
- **Search, NLP & Gen AI:** Apache Lucene, OpenNLP, Retrieval-Augmented Generation (RAG), Tool Calling, Agentic Workflows, LLM Orchestration, LLM APIs (OpenAI, Qwen, Llama)
- **System Design:** Distributed Systems, System Design, Data Structures & Algorithms

Experience

Zoho Corporation

Member Technical Staff

Chennai, India (On-site)

June 2023 - Present

- **Java Backend & REST APIs:** Designed and developed scalable Java and Spring Boot backend services and REST APIs powering an enterprise AI assistant integrated across multiple Zoho SaaS products in high-traffic production workloads.
- **Microservices & Distributed Systems:** Built and maintained Spring Boot microservices for agent orchestration, search, and multi-vendor LLM routing with fault-tolerant fallback chains, feature-flag rollouts, and service-to-service communication.
- **Multithreading, Persistence & Caching:** Implemented multithreaded components with custom thread-pool executors for parallel task dispatch; designed normalized MySQL schemas with indexed query optimization; integrated Redis Cluster (consistent hashing) for low-latency session and conversation caching.
- **Event-Driven Architecture & Search Relevance:** Designed asynchronous Apache Kafka producer-consumer pipelines for decoupled, scalable background processing; contributed scoring and ranking improvements to the Lucene- and OpenNLP-based search engine and optimized Trie-based Redis lookups.
- **Gen AI Engineering:** Built an end-to-end multi-query Retrieval-Augmented Generation (RAG) pipeline with streaming output; engineered prompt design, tool-calling framework, and agentic workflows powering production LLM use cases across OpenAI, Qwen, and Llama models.
- **Code Reviews & Mentoring:** Reviewed pull requests and mentored junior developers on Java backend design, multithreaded programming, async pipelines, and LLM integration best practices.

Projects

Async Query Processing Service

- Built a Spring Boot REST service that ingests user queries and publishes them to Kafka for asynchronous processing by worker consumers; persisted state and history in MySQL via Spring Data JPA / Hibernate and added Redis caching for hot results.
- Implemented status-tracking REST APIs, idempotent producers, and resilient error-handling for failed messages.
- **Stack:** Java, Spring Boot, Spring Data JPA, Hibernate, Kafka, MySQL, Redis, Maven, REST APIs.

Education

SASTRA University

MTech, Artificial Intelligence & Data Science

Thanjavur, Tamil Nadu

Aug 2021 - May 2023

JNTUH College of Engineering, Manthani

BTech, Computer Science

Manthani, Telangana

Jun 2017 - Jul 2021