Chitransh Saxena

SENIOR STAFF SOFTWARE ENGINEER — GENAI & DISTRIBUTED SYSTEMS

chitransh033@gmail.com | linkedin.com/in/chitransh22 | github.com/Chitransh-Saxena | leetcode.com/u/chitransh22 | chitransh.pulsar-projects.org

Senior engineer with **7+ years building and scaling distributed systems** — the last **3+ in production GenAI** (RAG, embedding pipelines, synthetic data generation) on enterprise-scale infrastructure. **Avid Claude Code user (1.5+ years)**, shipping rapid AI-assisted POCs across enterprise and personal projects.

TECHNICAL SKILLS

Languages: Go, Java, Python, C++, SQL, NoSQL

GenAI & ML: Retrieval-Augmented Generation (RAG), Synthetic Data Generation, LLM Fine-tuning & Extended Pre-training, Embedding Models, Semantic & Hybrid Search, Dense/Sparse Retrieval (BM25, SPLADE), Re-ranking, Multimodal Ingestion, Prompt Engineering, Agentic Coding (Claude Code), Model Context Protocol (MCP)

Search & Retrieval: MilvusDB, Elasticsearch, OpenSearch, Faiss, ANN Indexing (HNSW, IVF, Quantization), Hybrid Retrieval

Databases & Streaming: PostgreSQL, MS SQL Server, MongoDB; Apache Kafka (Exactly-Once, Consumer Groups, Partitioning, Offset Management), Redis Streams, Dead Letter Queues

Distributed Systems: Event-Driven Architecture, Stream Processing, Microservices, High Availability, Observability, SRE Practices, Design Patterns, TDD, CI/CD

Frameworks & Infra: Spring Boot, Resilience4j, gRPC, REST, Node.js/React; Kubernetes, Docker, Helm, Azure, Grafana, Prometheus, Jaeger

PROFESSIONAL EXPERIENCE

IBM India Software Labs
Aug 2024 – Present · Bangalore

Senior Staff Software Engineer

watsonx Code Assistant for Z — configurable SaaS embedding & RAG platform (Go codebase)

- Architected & led (team of 4 devs + 2 QEs) an end-to-end embedding + RAG pipeline processing **8K+ code pairs**, 50+ documents (500+ pages each) and 250+ runtime examples — multimodal, across multiple embedding models; scaled ingestion over 3 message-queue streams and 4 decoupled services.
- Led a BM25 study and integrated **SPLADE** for learned sparse retrieval → **Recall@20 +35%**. Prototyped vector search on **MilvusDB** (HNSW/IVF, quantization, <900ms inference), then ported to a hybrid **Elasticsearch (x86)** / **OpenSearch (s390x, IBM Z)** stack.
- Built Redis Streams infrastructure (consumer groups, DLQ, autoscaling, Grafana) at **sub-100ms / 99.95% delivery**; wrote a novel **LSP-based chunking** algorithm over 10K+ artifacts exposed via custom **MCP tools** (Python, MongoDB, GraphQL).
- Re-architected a monolith (Strategy pattern, 3+ language pairs) → **90% reuse, 100% throughput**; adopted by 4+ teams, cutting dev time 60% — **Star of the Month within 2 months**.
- Built a synthetic data-generation pipeline: **50K training pairs from a 5K base (10×)**, TDD-validated, to fuel LLM extended pre-training.

Walmart Global Tech India
Oct 2021 – Jul 2024 · Bangalore

Software Engineer 3

- Designed a fault-tolerant, event-driven stream-processing platform sustaining **1.5M transactions/day @ 50 TPS** on a 3-broker Kafka cluster (RF=3) — idempotent producers, **exactly-once semantics**, DLQ isolation, **99.99% availability SLO**.
- Cut **p99 consumer lag 2s → 0.5s**; collapsed 15+ consumer queries into 3; compressed the wire format **67% (150 → 50 bytes)**; eliminated 2 single points of failure.
- Engineered an extensible plugin framework handling **50+ event types across 15+ topics** (Strategy/Factory/Decorator with Resilience4j circuit breakers, bulkhead isolation, backoff retries).
- Ran 12+ autoscaled microservices on Kubernetes (Helm, HPA); built a MERN control plane automating releases **7–10 days → 45s** across 20+ repos; drove SRE-led P1 response → **MTTR –40%** (centralized logging + Jaeger tracing).

NielsenIQ
Jul 2019 – Oct 2021 · Chennai

Software Engineer

- Decomposed 3 monolithic services into a horizontally scalable microservices architecture with leader-election coordination → **throughput +60% (100 → 160 RPS)**, 2 SPOFs removed.
- Re-engineered latency-critical **C++ data pipelines → 55% runtime reduction (20s → 9s)** across 15K+ LOC via CPU/memory profiling, leak elimination and hot-path restructuring; owned on-call rotations.

EDUCATION

B.Tech, Information Technology — 88%
SRM Institute of Science and Technology · 2015 – 2019