Saurav Sharma

+91-7678582974 | sauravsharma3001@gmail.com | github.com/saurav470 | linkedin.com/in/saurav-kumar-sharma-sks

SUMMARY

Full-Stack Engineer with 3 years building high-throughput distributed systems in **Python (FastAPI, Django)**, **Node.js (Express, NestJS)**, **PostgreSQL**, and **AWS/Azure**, with **React/Next.js (TypeScript)** frontends. Designed microservices, async job queues, and real-time infrastructure handling **50K+ req/day** at **99.9% uptime**. Experienced with event-driven architectures (**Kafka**, **Redis Pub/Sub**), container orchestration (**Docker**, **Kubernetes**), and shipped **GenAI** features (**RAG pipelines**, **AI agents**, **MCP tool-calling**) into production.

EDUCATION

NSUT East Campus, Delhi | *B.Tech in ECE* CGPA: 8.3 / 10 Aug 2019 - Jul 2023

WORK EXPERIENCE

Devkraft | *Software Development Engineer* | Gurugram, India Jun 2024 – Present

- Built and scaled **Python microservices** (**FastAPI**, **Django REST Framework**) handling **50K+ requests/day** across 20+ services using async I/O and Redis-backed caching.
- Developed **Node.js** REST APIs with **Express** and **NestJS**, implementing JWT authentication and role-based access control.
- Built **React/Next.js** applications with **TypeScript** and **Redux**, implementing **SSR/SSG** patterns serving **50,000+ daily users** with sub-second page loads.
- **Led Engineer real-time WebSocket infrastructure** (asyncio, Redis Pub/Sub) supporting **10K+ concurrent connections** with persistent messaging and low-latency delivery.
- Designed **distributed task queues** (Celery, RQ) and **Kafka** event pipelines processing **500K+ messages/hour** with fault-tolerant delivery.
- Reduced **database latency by 65%** by optimizing PostgreSQL and MongoDB queries through indexing, Django ORM N+1 fixes, and schema improvements.
- Implemented **HLS adaptive bitrate streaming** (FFmpeg, AWS CloudFront), reducing buffering by **60%** and improving video playback performance.
- Built **LLM-powered RAG pipelines** (LangChain, pgvector) and **agentic workflows** (LangGraph) with **MCP tool-calling** and multi-provider LLMs (OpenAI, Claude), reducing inference costs by **40%** through prompt optimization and caching.
- Deployed cloud-native services on **Docker** and **Kubernetes** (auto-scaling, rolling updates), achieving **99.9% uptime** and **70% faster** incident recovery.

Alyssum Global Services | *Software Engineer Intern* | Remote Aug 2023 – Mar 2024

- Built **REST APIs** (**FastAPI**, **Node.js/Express**) on AWS Lambda and EC2 with JWT authentication and input validation, improving response time by **45%**.
- Automated **CI/CD pipelines** using GitHub Actions and Docker, reducing deployment cycles by **40%** across environments.

SKILLS

- **Languages:** Python, TypeScript, JavaScript, Go, SQL, Bash
- **Backend:** FastAPI, Django, Django REST Framework, Node.js, Express.js, NestJS, REST APIs, GraphQL, gRPC, WebSockets, asyncio, Celery, RQ, SQLAlchemy, Microservices
- **Frontend:** React.js, Next.js (SSR/SSG), Redux, TypeScript, HTML5, CSS3
- **Databases & Messaging:** PostgreSQL, MongoDB, Redis, DynamoDB, Apache Kafka, RabbitMQ, Amazon SQS/SNS
- **Cloud & DevOps:** AWS (Lambda, ECS, EC2, S3, CloudFront, CDK), Azure, Docker, Kubernetes, GitHub Actions, CI/CD
- **GenAI & LLMs:** OpenAI API, Claude (Anthropic), Gemini, AWS Bedrock, Hugging Face, Prompt Engineering, Function/Tool Calling, Structured Outputs, Embeddings, Fine-tuning (LoRA)
- **Agentic AI & RAG:** LangChain, LangGraph, LlamaIndex, CrewAI, AutoGen, MCP Servers, AI Agents, RAG Pipelines, Hybrid Search (BM25 + Vector), Reranking, pgvector, FAISS, Pinecone, Weaviate, Qdrant, LangSmith, Langfuse, Evals & Guardrails
- **Testing, Observability & Tools:** Pytest, Unittest, Locust, OpenTelemetry, Prometheus, Grafana, AWS CloudWatch, Distributed Tracing, FluentD, Git, Claude Code

PROJECTS

Medical AI Platform — Real-Time Diagnostic Assistant Devkraft

- Architected an **agentic AI platform** with API gateway, circuit breakers, and health checks, orchestrating **LLM inference**, **AI agents (LangGraph)**, and **MCP tool servers** for real-time diagnostic workflows.
- Built **doctor and patient portals** in **Next.js** and **TypeScript** (AI chat over past consultations and prescriptions, video-call UI) on **Django REST** and **FastAPI** services.
- Built a **patient-record RAG pipeline** over consultation transcripts, prescriptions, and video-call clips with **hybrid search** (BM25 + vector) and reranking, reducing hallucinations in patient chat.

- Deployed **WebRTC** video consultations with a Go signaling server and doctor handoff under **200ms**; built FluentD + Grafana dashboards for LLM cost and SLA tracking, reducing debugging time by **40%**.

Transcendify — Video Transcoding Platform Personal

- Created a **distributed video processing queue** with Python workers, Celery, and Redis handling **1,000+ videos/day** with **FFmpeg HLS** transcoding.
- Delivered **multi-resolution output** (1080p/720p/480p) via AWS S3 and CloudFront CDN with **sub-500ms** enqueue latency.

ACHIEVEMENTS

- **Microsoft Ideathon'21:** India Semifinalist, **Top 8** nationally. Sep 2021 – Feb 2022
- **HackNSUT: Top 10** among 300+ participants in a competitive coding hackathon. Apr 2022
- **HackTheLeague (by IIT): Top 50** of 1000+ applicants, Pan-India tech innovation challenge. Jun 2022