Roshan Yadav

Mumbai, India | +91 9834905805 | roshany.ir.ry@gmail.com | linkedin.com/in/roshanyadav026 | github.com/crazycodeop

Summary

Software Engineer with **2+ years** of experience architecting highly scalable, cloud-native B2B SaaS platforms. Specialized in backend distributed systems (**Python, Django, FastAPI**), **AWS Cloud Infrastructure**, and integrating enterprise-grade **GenAI / RAG architectures**. Proven track record of translating complex client requirements into technical roadmaps, optimizing high-throughput databases, and delivering robust, fault-tolerant applications that drive operational efficiency and client satisfaction.

Skills

Languages & Core: Python, JavaScript, SQL (PostgreSQL, MySQL), HTML5/CSS3
Backend & Architecture: Django, Django REST Framework, FastAPI, Celery, RESTful Microservices, System Design
Cloud & DevOps: AWS (EC2, S3, RDS, IAM), Docker, Linux, Jenkins CI/CD, Nginx, Serverless DBs
AI & Machine Learning: Gemini API, LangChain, LangGraph, RAG, Vector Databases, Prompt Orchestration
Frontend & Integration: React, Third-Party API Integration (Payment Gateways, eSignature), OAuth2, JWT
Practices & Performance: High-Availability Scaling, Redis Caching, ORM Tuning, Automated & Performance Testing

Experience

VakilDesk Technologies Pvt. Ltd. Remote
Software Engineer Oct 2024 – Oct 2026

- Spearheaded end-to-end architecture and client requirement gathering for **4 enterprise B2B SaaS platforms** using **React and Django**, achieving **100% on-time delivery** across **10+ corporate clients**.
- Engineered highly available cloud infrastructure on **AWS (EC2, S3, RDS)** with **Docker** and **Jenkins CI/CD**, cutting release deployment cycles by **35%** while maintaining **99.9% SLA uptime**.
- Pioneered production **GenAI RAG pipelines** using **Gemini API**, **LangChain**, and vector embeddings for semantic retrieval, reducing automated legal drafting turnaround time by **60%**.
- Orchestrated fault-tolerant microservice integrations across **DocuSign, eMudhra, GST, and WhiteBooks APIs** via **Celery**, processing **10,000+ monthly asynchronous transactions** with zero data loss.
- Optimized high-throughput **PostgreSQL** databases by refactoring complex Django ORM queries and adding **Redis** caching layers, eliminating N+1 bottlenecks and slashing API latency by **40%**.
- Reduced production cloud hosting expenses by **50%** by resolving memory leaks and rightsizing **AWS EC2** instances from 32 GiB to 8 GiB RAM without performance degradation.
- Developed real-time document collaboration workflows utilizing **WebDAV** for live MS Word editing, automated version history, and PDF rendering across **5,000+ legal documents**.
- Established automated quality assurance and load testing suites using **pytest**, achieving **85%+ test coverage** and cutting post-deployment production regressions by **45%**.
- Strengthened multi-tenant application security across **10+ client environments** by enforcing **JWT, Google OAuth2, RBAC, CORS, and rate limiting** to protect sensitive legal payloads.

Projects

Docvyn – Enterprise Document Intelligence Platform | *FastAPI, React 18, pgvector, Celery, Groq* | GitHub

- Architected a multi-tenant RAG platform using **FastAPI, SQLAlchemy 2.0, and Supabase PostgreSQL (pgvector)**, indexing **384-dimensional FastEmbed** vectors with page-level source citations.
- Integrated non-blocking **Celery and Redis** ingestion workers with **Groq (Llama 3.3 70B)**, **Gemini** fallback, and prompt-injection defense to deliver **sub-second** grounded Q&A at zero embedding API cost.
- Validated end-to-end multi-tenant SQL vector isolation, JWT token rotation, and React 18 UI reliability across **48 automated unit and integration tests** using **Pytest, HTTPX, and Vitest**.

FlowForge AI – Visual AI Workflow Orchestrator | *FastAPI, React Flow, Celery, Upstash Redis, Supabase* | GitHub

- Built a dual-portal DAG workflow platform using **FastAPI, React Flow, and PostgreSQL 17**, achieving **sub-40ms** asynchronous task dispatch via **Upstash Redis** and distributed **Celery** workers.
- Implemented a topological DAG scheduler (**Kahn's algorithm**) with distributed worker heartbeats, zombie lease recovery, and **SHA-256** idempotency to guarantee **100%** durable state execution.
- Designed deterministic LLM guardrails over **Groq LPUs and OpenAI** with **Pydantic** schema repair, role-gated human approval queues, and **62 Pytest** tests at a **100% pass rate**.

Education

Thadomal Shahani Engineering College Mumbai, India
B.E., Electronics & Telecommunications Engineering (CGPA: 8.82 / 10.00) June 2024

- Relevant Coursework: Data Structures, Analysis of Algorithms, Object-Oriented Programming, Database Management Systems (DBMS)