

SAIRAM UGGE

+91-6300220467 ◊ Hyderabad, Telangana, India
uggesairam0000@gmail.com ◊ LinkedIn ◊ GitHub

Work Experience

Software Engineer — Backend & GenAI Platforms
Ascendion, India

Jul 2024 – Present

- Co-built and production-hardened **Pensieve**, an AI process-orchestration engine driving **multi-agent LLM workflows** through human-in-the-loop approval gates and real-time streaming, with **governed LLM routing** across cloud providers — adopted by **2K+ users daily** across different domains
- Co-built the backend powering **AAVA Code**, an AI coding plugin for VS Code, re-architecting it from a single-agent prototype into a **multi-agent orchestration system** (Main-Agent + sub-agent using **crewAI flows**, **150+ skills**, **40+ tools**, **~60 commands**) — adopted by **3K+ users daily** across **5+ client environments**
- Co-built a GenAI-powered wireframe generator that converts product inputs (PRDs, sketches, prompts) into production-ready UI artifacts, reducing UX iteration cycles by **40%** (5 review rounds to 3) and accelerating prototyping by **60%**, serving **500+ users daily**
- Co-built a prompt-to-React code generation system that converts generated wireframes into production-ready modular components and routing logic, cutting frontend development time by **50%** (2 days to 1 per feature) and manual coding effort by **55%**, adopted by **2K+ developers** across **50+ teams**
- Co-built a prompt-driven execution engine using LLM-based agent orchestration with RAG and ReAct pipelines, compressing workflow planning from **1.5 hours to 15 minutes** and raising first-pass acceptance from **65% to 85%**, enabling automated generation of structured product artifacts for **1.5K+ users**
- Owned **backend systems end-to-end** (API design, data modeling, event-driven pipelines), scaling throughput **10×** at **sub-150ms** latency via a decoupled pub/sub architecture (**Redis Streams + SSE**) that eliminated polling to process **10K+ events/day** for **2.5K+ users**

Skills

Programming Languages	Python, Go
Backend & Distributed Systems	FastAPI, gRPC, REST, GraphQL, Microservices, Event-Driven
Data & Storage	PostgreSQL, DynamoDB, MongoDB, Redis, ChromaDB
Real-Time & Messaging	Apache Kafka, Redis Streams, WebSockets, SSE
GenAI Systems	LLM Agent Orchestration, RAG, Prompt Engineering, Vector DBs & Embeddings
Cloud & DevOps	AWS, Docker, Kubernetes, CI/CD (GitHub Actions, Azure)
Observability	OpenTelemetry, Prometheus, Grafana
Frontend	React, Angular, Next.js, Tailwind CSS
Developer Tools	Git, GitHub, Linux, Nginx

Achievements

- Secured top global ranks in competitive programming contests, including **Google Code Jam 2023** (AIR 420; 3,687 / 85,000+) and **Meta Hacker Cup 2022** (4,048 / 70,000+)
- Reached the top tier in **Flipkart Grid 2022** (1,325 / 40,000+)
- Mentored **250–300** students in **Data Structures and Algorithms (DSA)** through the Smart Interviews program

Education

B.Tech, Electronics & Communication Engineering, VNR VJIET — CGPA **9.08**

2021 – 2024

Projects

OPS - gRPC Microservices



- Architected a polyglot (**Go + Python/FastAPI**) **event-driven microservices** system with gRPC internal RPC and a REST gateway, featuring **etcd-backed leader election**, **ACID inventory reservations**, a **Saga orchestrator**, and metric-based read routing across **PostgreSQL replicas**
- Engineered reliability tooling — **Debezium/WAL CDC** outbox, Bloom filters, **Redis caching** (cache-aside, single-flight dedup), rate limiting, DLQ + idempotency, and **OpenTelemetry/Prometheus/Grafana** observability via **Envoy L7**