

Sourav Rajvi

souravrajvi@gmail.com | +91-6377278247 | [LinkedIn](#) | [GitHub](#)

EXPERIENCE

Software Engineer

March 2026 – Present

Planys Technologies

- Developed an **event-driven ROV telemetry backend** using **Java, Spring Boot, Apache Kafka, and PostgreSQL**, processing **500+ telemetry events/sec across 50+ simulated vehicles** with **sub-200 ms p95 ingestion latency**.
- Engineered **Kafka producer-consumer pipelines** with **vehicle-based partitioning, idempotent processing, retries, and dead-letter handling**, supporting **1M+ telemetry events per test run** while preventing duplicate persistence during consumer retries.
- Built **20+ REST APIs** using **Spring MVC, Spring Security, Spring Data JPA, and Hibernate** for **vehicle management, mission lifecycle, device configuration, telemetry history, and alert workflows** with role-based access control.
- Implemented **Redis-backed live vehicle state and WebSocket streaming**, reducing latest-state retrieval latency by **~60%** and delivering near-real-time telemetry and operational alerts to monitoring clients.
- Developed a configurable **real-time alerting engine** for **battery, depth, connectivity, and sensor-health events**, using threshold rules and hysteresis to reduce repeated alert generation by **70%+** during sustained fault conditions.
- Optimized telemetry persistence using **PostgreSQL time-based partitioning, composite indexing, and batch inserts**, improving historical telemetry query performance by **~45%**, with **Docker, JUnit/Testcontainers, CI/CD, and AWS CloudWatch** for deployment and monitoring.

Software Engineer

May 2025 – Feb 2026

AGSuite Technologies

- Architected an **asynchronous financial middleware platform** using **Java, Spring Boot, Spring MVC, and REST APIs**, integrating **NetSuite with Zoho Books** and processing **10,000+ vendor invoices daily** with **99.9% uptime**.
- Engineered **event-driven invoice processing** using **AWS SQS, S3, EC2, and Auto Scaling**, handling month-end bursts of **5,000+ PDFs/hour** while reducing batch processing time from **2+ hours to ~20 minutes**.
- Developed persistence and transaction workflows using **Spring Data JPA, Hibernate, and Amazon RDS PostgreSQL**, implementing **ACID transactions, JSONB audit records, and database-level idempotency** to prevent duplicate financial transactions.
- Designed resilient processing with **SQS visibility timeouts, retry handling, Dead Letter Queues, and CloudWatch alarms**, achieving **99.99% processing success** and reducing production failure detection to **under 2 minutes**.
- Containerized and deployed **Spring Boot services** using **Docker, Maven, AWS EC2, Auto Scaling, IAM, and VPC**, with automated **CI/CD** and workload-based scaling that reduced infrastructure costs by **~60%**.

Software Developer Intern

Aug 2024 – Jan 2025

Amazon

- Optimized high-scale **Payment Contract backend** using **Java, Spring Boot, and Python**, reducing load time from **29s to 4s** across Amazon business workflows and earning an internal accolade for performance engineering excellence.
- Resolved a critical production regression using **Python monitoring automation** and **AWS CloudWatch**, achieving **100% rollback success** and reducing errors by **17%** across microservices serving **millions of users**.

TECHNICAL SKILLS

Languages: Java, Python, C++, SQL

Backend Technologies: Spring Boot, Spring MVC, Spring Security, Spring Data JPA, Hibernate, REST APIs, Microservices, Apache Kafka, RabbitMQ, Swagger/OpenAPI, FastAPI

Databases & Search: PostgreSQL, MySQL, Redis, MongoDB, DynamoDB, pgvector, Pinecone

Cloud & Infrastructure: AWS (EC2, Lambda, S3, SQS, RDS, CloudWatch, Auto Scaling, IAM, VPC), Docker, Kubernetes

Build & DevOps: Git, GitHub, Bitbucket, GitHub Actions, Jenkins, CI/CD, Maven, Gradle, Agile/Scrum

GenAI & LLM: Spring AI, LangChain, OpenAI API, RAG Pipelines, Vector Embeddings, Vector Search, LLMs (LLaMA, GPT)

ACHIEVEMENTS & OPEN SOURCE

- **Core Contributor to Unsloth AI** (69K+ stars, YC S24), contributing to GPU/installer infrastructure, model loading, and agent tooling.
- Merged **50+ PRs in 2026** across other YC-backed, high-star open-source projects — Dify (142K+ stars), AnythingLLM (63K+ stars), LiteLLM (56K+ stars), and Mastra (16K+ stars) — spanning API security and workflow UX.
- Built and open-sourced **OpenInference CLI**, a hardware-aware package manager for local LLMs that profiles CPU/GPU/RAM and recommends compatible models from a catalog of **150+ models**; reached **11,000+ npm downloads**.
- **Achieved a peak LeetCode rating of 2101** (top 6% globally) and secured a **best contest rank of 39** in Weekly Contest 494 (22 Mar 2026).

EDUCATION

Bangalore Institute of Technology

Bachelor of Engineering in Electronics and Communication Engineering

2024