

HIMANSHU BALIYAN

Software Engineer — Backend & AI Systems
+91 70068 02968 · himanshubaliyan4000@gmail.com · linkedin.com/in/himanshubaliyan

SUMMARY

Software engineer with 12 months of experience building production backend services in Java and Spring Boot and shipping LLM-powered features including RAG pipelines and agentic workflows.

EXPERIENCE

Software Engineering Intern Jan 2025 – Jan 2026
Airtel Digital Gurugram, India

- Built and maintained 6 backend microservices and 25+ REST endpoints using Java, Spring Boot and PostgreSQL, supporting enterprise analytics workflows used by 4 internal teams.
- Developed AI-powered backend workflows using LLM APIs, LangChain and a RAG retrieval layer, cutting enterprise document lookup from ~2 hours to under 5 minutes per request.
- Designed distributed Apache Spark and Kafka pipelines ingesting ~12M telemetry events per day, reducing batch processing time by 35% and sustaining 99.5% pipeline uptime.
- Instrumented 8 services with Prometheus, Grafana, Loki, Tempo and OTLP tracing, reducing mean time to diagnose production incidents by ~40% and raising monitoring coverage to 90%.
- Automated containerised deployments with Docker, Kubernetes, Git and CI/CD pipelines, cutting release time from ~45 minutes to under 10 and reducing failed deployments by 60%.
- Optimised SQL queries and service APIs, improving average p95 response time from ~850 ms to ~300 ms under production load.

SKILLS

Programming Languages	Java, Python, SQL, JavaScript, Bash
Libraries / Frameworks	Spring Boot, REST APIs, Microservices, Apache Spark, Apache Kafka, LangChain, LangGraph
Databases	PostgreSQL, MongoDB, OracleDB
AI & LLMs	RAG Pipelines, LLM API Integration, Agentic Workflows, Tool Calling, MCP, Prompt Engineering, Embeddings & Vector Search
Cloud & DevOps	AWS, Docker, Kubernetes, Linux, Git, CI/CD, Prometheus, Grafana, Loki, Tempo, OpenTelemetry

PROJECTS

- Agentic Job Automation Platform** | *Python, LangGraph, LLM APIs, Tool Calling*
Multi-agent workflow platform using graph-based orchestration, tool calling and structured memory to plan and execute job application tasks end to end, with state persistence between agent steps.
- Enterprise RAG Platform** | *Python, LangChain, Vector Store, REST API*
Retrieval-Augmented Generation service for semantic search over enterprise documents: chunking, embedding, retrieval and grounded answer generation with source citations, exposed as a REST API.
- Superset AI — Natural Language to SQL Assistant** | *Python, LLM APIs, Apache Superset, SQL*
Added LLM-powered querying to Apache Superset, generating schema-aware SQL from plain English with validation before execution so analysts could self-serve without writing queries.
- IoT DataHub — Distributed Spark Pipeline** | *Apache Spark, Kafka, Java, Docker*
Fault-tolerant streaming pipeline ingesting high-volume IoT telemetry through Kafka into Spark for transformation and aggregation, with checkpointing and recovery on failure.

EDUCATION

Master of Computer Applications — CGPA 8.4 Sep 2022 – Sep 2024
Hindustan University Chennai, India