Harsha Vardhan Yellela

United States | +1-248-497-9965 | harsha.yellela@gmail.com | har5ha.in | LinkedIn | GitHub

Summary

Software Engineer with 3+ years of experience building scalable, cloud-native applications using Java, Python, and modern frameworks like Spring Boot, FastAPI, and Node.js. Experienced in designing microservices, REST/GraphQL APIs, and event-driven systems with Kafka, along with deploying solutions on AWS using Docker and Kubernetes. Hands-on expertise in AI/LLM technologies, including RAG pipelines, LangChain, and Bedrock, to deliver intelligent, data-driven systems. Strong background in performance optimization, CI/CD automation, and observability, focused on delivering reliable, secure, and high-impact solutions in fintech and enterprise domains.

Skills

- **Programming Languages:** Java, Python, C++, JavaScript, TypeScript, SQL
- **Backend Development:** Spring Boot, FastAPI, Node.js, Express.js, RESTful APIs, GraphQL APIs, Microservices Architecture, Distributed Systems, API Gateway, OAuth2/JWT Security
- **Frontend Development:** React.js, JavaScript (ES6+), Component-Based Architecture, Responsive UI Development
- **AI/ML & LLM Technologies:** Agentic AI, RAG Pipelines, LangChain, CrewAI, AWS Bedrock, Semantic Search, Prompt Engineering, NLP Workflows
- **Cloud & DevOps:** AWS (EKS, Lambda, S3, API Gateway, Fargate, Bedrock, OpenSearch), Docker, Kubernetes, Terraform, CI/CD Pipelines, GitHub Actions, Jenkins, Infrastructure as Code (IaC)
- **Event Streaming & Messaging:** Apache Kafka, Event-Driven Architecture, Real-Time Data Processing
- **Databases & Caching:** MySQL, PostgreSQL, DynamoDB, MongoDB, Redis, OpenSearch, Query Optimization, Data Modeling
- **System Design & Architecture:** Scalable System Design, High Availability Systems, Fault Tolerance, Performance Optimization, Multi-Tenant Architecture
- **Monitoring & Observability:** ELK Stack (Elasticsearch, Logstash, Kibana), Prometheus, Grafana, Logging, Metrics, Alerting, MTTR Reduction
- **Software Engineering Practices:** Agile/Scrum, Test-Driven Development (TDD), Unit & Integration Testing, Debugging, Code Reviews, CI/CD Automation, Performance Tuning
- **Tools & Frameworks:** Git, GitHub, REST Clients, GraphQL, Async Processing, Multithreading, Containerization

Experience

Software Engineer – PNC Bank, USA **Jan 2026 – Present**

- Engineered scalable microservices using Java (Spring Boot) and Python, to modernize fraud detection workflows, to reduce transaction anomaly detection latency by 35%, for risk analytics and compliance teams across retail banking systems.
- Architected an AI-driven RAG pipeline using LangChain and AWS Bedrock, to enable contextual financial insights from structured and unstructured data, to improve decision accuracy in fraud investigations, for internal fraud analysts and audit stakeholders.
- Implemented event-driven data pipelines using Apache Kafka and Redis caching, to process high-volume transaction streams in near real-time, to increase system throughput and reduce API response times below 200ms, for customer-facing banking applications.
- Containerized and deployed services on AWS EKS (Kubernetes + Fargate), to ensure high availability and auto-scaling, to support peak loads of 50K+ daily transactions, for enterprise banking platforms.
- Reduced production downtime by improving observability with ELK stack, Prometheus, and Grafana, to proactively identify system anomalies, to lower MTTR by 40%, for DevOps and SRE teams.
- Developed secure REST and GraphQL APIs with Node.js and Spring Security (OAuth2/JWT), to enforce fine-grained access control, to ensure compliance with financial security standards, for internal and external API consumers.
- Optimized CI/CD pipelines using Jenkins and GitHub Actions, to automate build, test, and deployment workflows, to accelerate release cycles by 30%, for cross-functional engineering teams.

Software Engineer Research Assistant – Agentic AI | Lawrence Technological University, USA **Jan 2025 – Dec 2025**

- Designed multi-agent orchestration systems using CrewAI and LangChain, to automate research workflows like literature review and data extraction, to reduce manual effort by 70%, for university research teams.
- Built and deployed 3 persistent AI agent services on AWS EKS with Fargate, to enable scalable execution of agent pipelines, to support semantic search across 10K+ documents with sub-second latency, for academic and research applications.
- Developed FastAPI and GraphQL-based backend services integrated with AWS Bedrock, to expose LLM-powered capabilities, to deliver dynamic query responses and summaries, for AI-driven research platforms.
- Implemented OpenSearch Serverless and Redis caching layers, to optimize document retrieval and session management, to enhance system responsiveness and reduce redundant computations, for high-frequency research queries.
- Automated infrastructure provisioning using Terraform and GitHub Actions CI/CD, to streamline environment setup, to reduce provisioning time from 2 hours to 15 minutes, for cloud-native deployments.
- Conducted benchmarking between no-code (n8n) and coded (CrewAI) agent frameworks, to evaluate performance trade-offs, to guide architectural decisions for future AI research initiatives, for faculty and engineering stakeholders.
- Built scalable backend pipelines in Python with asynchronous processing, to handle concurrent agent tasks, to improve execution efficiency and system throughput, for multi-agent AI workloads.

SDE-1 (LN Technical Consultant) | Infor India Pvt. Ltd., Hyderabad, India **Apr 2022 – Dec 2023**

- Developed enterprise-grade RESTful APIs using Java and Spring Boot, to integrate ERP systems with external clients (Ferrari, Boeing, Triumph), to process 500+ daily transactions reliably, for global manufacturing and supply chain systems.
- Engineered serverless microservices using AWS Lambda, API Gateway, and S3 triggers, to automate data ingestion and event processing, to improve system scalability and reduce infrastructure overhead, for enterprise clients.
- Diagnosed and resolved 15+ critical data pipeline issues across MySQL and distributed systems, to stabilize batch processing workflows, to reduce failure frequency from weekly to monthly, for production operations teams.
- Built backend services in C++ for performance-critical modules, to optimize data processing and reduce latency in ERP integrations, to enhance system efficiency for high-volume enterprise workloads.
- Containerized applications using Docker and implemented CI/CD pipelines, to standardize deployments across environments, to reduce release turnaround time by 25%, for multi-client delivery teams.
- Designed frontend components using React.js and JavaScript, to improve user interaction with ERP dashboards, to enhance usability and reporting efficiency for business users.
- Collaborated in Agile teams to deliver features across microservices and distributed architectures, to ensure timely delivery and system reliability, for enterprise stakeholders and client integrations.

Education

Master of Science in Computer Science · GPA: 3.77/4.0 **Jan 2024 – Dec 2025**

Lawrence Technological University, Southfield, MI