

SUDHEENDRA SABNAVISU

+1 857-693-9339 | sudheendra2108@gmail.com | sudheendra21.live | linkedin.com/in/sudheendra21 | Boston, MA

SUMMARY

Backend Software Engineer with 4+ years of experience designing scalable APIs, microservices, and distributed systems in Java, Spring Boot, Python, and FastAPI. Strong in Kafka-based event-driven architecture, SQL and data modeling, AWS, and CI/CD, with hands-on experience building GenAI features (RAG pipelines, vector search, LLM integrations). Proven record of shipping reliable, low-latency backend services in production.

TECHNICAL SKILLS

Languages: Java, Python, JavaScript, SQL, Bash

Backend: Spring Boot, FastAPI, REST APIs, Microservices, API Design, System Design, Service Integration, OAuth 2.0 / RBAC

Distributed Systems: Kafka, RabbitMQ, Redis, Event-Driven Architecture

Databases & Data: PostgreSQL, PGVector, DynamoDB, Elasticsearch, SQL Query Optimization

Cloud & Infrastructure: AWS, GCP, Azure, Docker, Kubernetes, Terraform, Packer, GitHub Actions, Jenkins

AI and LLMs: RAG, LangChain, Embeddings, Semantic Search, MCP

PROFESSIONAL EXPERIENCE

Banner Life Family of Companies

Sep 2025 – Present

Software Engineer (Contract)

Remote

- Owned end-to-end design, development, and production deployment of 2 high-throughput REST APIs, supporting peak loads of 1,000 QPS to synchronize lead, policy, and application-status data across CRM and underwriting platforms.
- Secured 100% of external API traffic using OAuth 2.0 Client Credentials Flow, token-based authentication, and RBAC, while implementing structured logging, payload validation, and automated recovery with up to 5 retry attempts.
- Architected and delivered a Python-based Generative AI platform for an enterprise learning and compliance system, launching 3 production capabilities: RAG-powered assistance, automated assessment generation, and AI-driven course summarization.
- Built a scalable RAG pipeline using Python, LLM APIs, embeddings, semantic search, metadata filtering, and citation-backed responses, reducing learner content-discovery time by 60% and grounding responses in authorized course material.
- Automated generation of quizzes, answer keys, and difficulty-tagged assessments, reducing content-authoring time by over 80% and enabling instructors to generate a 10-question assessment in under 5 minutes with human review and approval.

Physis Investment

Jan 2025 – Apr 2025

Software Engineer

Boston, MA

- Built and deployed scalable FastAPI services for ImpactChat, a proprietary ESG research assistant using RAG, PostgreSQL, and PGVector to analyze 170M+ data points across 14K+ companies and 335K+ funds.
- Designed a Kafka-backed feedback pipeline to asynchronously capture thumbs-up/down ratings and user feedback for chatbot responses, enabling scalable analytics, model evaluation, and end-to-end traceability across prompts, responses, and LLM inference metadata.
- Developed streaming APIs and D3.js visualizations for portfolio analysis, industry benchmarking, and historical trends, delivering complex ESG insights through interactive charts, tables, and investor-facing dashboards.
- Built CI/CD pipelines using GitHub Actions and Terraform to automate build, testing, and deployment of FastAPI services to auto-scaled AWS EC2 infrastructure, reducing production deployment time by 65% and enabling weekly releases.

Salesforce

Jul 2021 – Aug 2023

Software Engineer

Bengaluru, India

- Migrated application workloads to AWS using Packer-based immutable images, VPC peering, private networking, and secure database connectivity, reducing infrastructure provisioning time by 70% and eliminating environment configuration drift.
- Built a Kafka-based onboarding event pipeline that consumed new-hire events from Salesforce and triggered SendGrid welcome emails, IT account provisioning, compliance training assignments, and manager notifications, decoupling 4 downstream workflows with retry and replay support.
- Designed and optimized scalable backend services and automation pipelines (Apex, JavaScript, batch processing), reducing system latency by 20% and improving data accuracy by 35%.
- Improved query performance through SQL optimization, achieving 25% faster execution on high-volume datasets.

Resideo

Jan 2021 – Jun 2021

Intern

Bengaluru, India

- Built an Elasticsearch indexing pipeline for thermostat telemetry and application logs, enabling searches by device ID, timestamp, temperature, setpoint, HVAC state, and error code; created Kibana dashboards that reduced production issue triage time by 40%.
- Developed a modular Java application generating customizable JSON payloads with object-oriented design patterns, enabling reuse across testing teams and improving testing efficiency by 30%.
- Authored API documentation following OpenAPI specifications with Swagger, standardizing interface contracts across services.

EDUCATION

Northeastern University

Boston, MA

Master of Science in Computer Software Engineering

Sep 2023 – Jul 2025

Amrita Vishwa Vidyapeetham

Amritapuri, KA

Bachelors in Computer Science

July 2017 – Jul 2021