

# Sai Nihal Konduti

## Software Engineer (AI)

Bloomington, IN | +1 (812) 369-9589 | [sainihalk2002@gmail.com](mailto:sainihalk2002@gmail.com) | [sainihalkonduti.com](http://sainihalkonduti.com) | [LinkedIn](#)

## SUMMARY

Backend & AI Engineer (Python/Java, Azure/GCP) building secure microservices and reliable AI systems (RAG, Agents). Scaled a multi-tenant platform for 100+ concurrent tenants, cutting latency by 98% using Azure ETL pipelines. Built Vertex AI workflows to process large datasets efficiently. Microsoft Certified: [Azure Fundamentals \(AZ-900\)](#) | [Azure AI Engineer Associate \(AI-102\)](#)

## EDUCATION

**M.S. in Computer Science** | Indiana University, **GPA:** 3.86/4 | Bloomington, IN Aug 2024 - May 2026

**B.Tech. in Computer Science** | B V Raju Institute of Technology, **GPA:** 8.75/10 | Hyderabad, India Aug 2020 - Jun 2024

## TECH STACK

**Languages:** Java | Python | TypeScript | SQL | C | JavaScript | R | HTML/CSS

**Architectures & Concepts:** Distributed Systems | Microservices | RESTful APIs | Postman | ETL Pipelines

**AI & Agentic Systems:** Generative AI | Large Language Models (LLMs) | RAG | LangChain | Prompt Engineering

**Frameworks & Libraries:** Spring Boot | Next.js | Node.js | FastAPI

**Databases & Data Platforms:** MySQL | Azure SQL | MongoDB | Redis | Tableau | Power BI

**Cloud & DevOps:** Azure | AWS | Google Cloud (Vertex AI) | Docker | Kubernetes | CI/CD (GitHub Actions)

**Methodologies:** Agile | Scrum

## EXPERIENCE

**Software Engineer (AI)** | Infoswift Corp. | Irvine, CA | Remote Sep 2025 – Dec 2025

- Built an asynchronous serverless ETL pipeline on Azure Functions using Docker to parallelize Gemini inference at scale, eliminating analytics bottlenecks for 20,000+ files and slashing processing latency by 98% (15m to 30s).
- Secured 100% data isolation for 100+ concurrent tenants and prevented cross-tenant leakage by engineering a multi-tenant microservice (FastAPI, Azure). Enforced tenant-scoped paths and HMAC-signed JWTs via GitHub Actions CI/CD.
- Reduced API error rates by 40% and maintained 99.9% uptime during traffic spikes by implementing distributed rate-limiting middleware (Redis, Token Bucket). Validated per-tenant quotas using TypeScript and Postman integration tests.
- Achieved 90% extraction accuracy for clinical data (vitals, labs), transforming unstructured reports into structured assets, by engineering a RAG evaluation framework (LangChain) to benchmark and optimize prompt strategies for Gemini Pro.

**AI Engineer** | Zion Cloud Solutions (ZionAI) | Glenview, IL | On-site Jun 2025 – Aug 2025

- Minimized review latency by 30% and enhanced Q&A reasoning by deploying distributed systems on Google Vertex AI (AutoML, Docker, K8s). Benchmarked few-shot prompting across 50+ repositories to streamline policy analysis.
- Slashed developer initialization time by 40% by architecting a Small Language Model (SLM) scaffolding (Node.js) to automate infrastructure setup. Developed a TypeScript and Next.js interface to streamline the developer onboarding.
- Enforced data integrity for 10000+ concurrent records by building a type-safe validation framework to enforce schema consistency across PostgreSQL and Azure SQL. Ensured high availability and compliance during Agile delivery sprints.

**Software Engineer (Cloud/AWS)** | Infoswift Corp. | Hyderabad, India | On-site May 2023 – Jun 2024

- Accelerated feature deployment cycles by 90% (reducing build times from 20 min to <2 min) by migrating manual tasks to a fully automated CI/CD pipeline using AWS CodePipeline and GitHub Actions for the production environment.
- Cut cloud infrastructure costs by 20% and improved incident response time by 50% by conducting EC2 rightsizing audits, automating S3 archival policies, and engineering an Amazon CloudWatch monitoring suite for 15+ servers.

## PROJECTS

**Hoosier Hub: Full Stack Social App** [[github.com/IU-HoosierHub/server](https://github.com/IU-HoosierHub/server)]

- Engineered a high-concurrency backend scalable to 50,000+ students and 1,000+ clubs, ensuring stability and sub-200ms latency, by architecting APIs (OpenAPI) and benchmarking MongoDB schemas.
- Established a production-grade DevOps environment by building a fully automated CI/CD pipeline with Docker and GitHub Actions, while securing the platform with JWT authentication using Spring Boot.

**Distributed Serverless Log Processing System with Azure** [<https://github.com/nihal2405/CloudLogProcessing>]

- Processed 10GB+ of daily raw logs in real-time by building a high-throughput ingestion engine using Azure Functions to parse, split, and structure unstructured data streams, eliminating 90% of manual parsing overhead.
- Reduced query time for system audits from hours to seconds by structuring unstructured text into queryable insights, using a parallel parsing engine to extract timestamps and severity levels into Azure SQL via optimized bulk inserts.