

NIKHIL KUMAR SAHU

Software Engineer — Backend, Cloud & Distributed Systems

☎ +91-6206910245

✉ heynikhil07@gmail.com

🌐 [linkedin.com/in/nikhilsahu07](https://www.linkedin.com/in/nikhilsahu07)

🐙 github.com/nikhilsahu07



nikhilsahu.dev

PROFESSIONAL EXPERIENCE

Software Engineer

Nov 2025 – Present

RaghuvarTech (Raghuvar IT Solutions LLP)

Ranchi, Jharkhand

- Own infra provisioning across dev/staging/prod using Terraform (VPC, EC2, RDS, IAM) as Infrastructure-as-Code, with GitHub Actions CI/CD pipelines enforcing plan-before-apply, drift detection, and SSM/OIDC-backed secrets management with zero long-lived credentials
- Architected **infra-ops**, a centralized multi-environment IaC platform with S3 remote state, DynamoDB state locking, reusable Terraform modules (networking/compute/storage/IAM), and Ansible roles for PostgreSQL/PostGIS, Nginx+Certbot, and PM2-based app deploys
- Built **lgtm-sidecar**, a production-grade, resource-capped observability stack (Grafana Alloy, Loki, Tempo, Prometheus, Pyroscope, Alertmanager) as a decoupled EC2 sidecar, with distributed tracing across microservices and rate-capped tail sampling for faster root-cause analysis
- Lead architecture reviews and cross-functional delivery across backend microservices, infrastructure, and RAG & Agentic AI systems; driving hands-on adoption of Kubernetes, Helm, ArgoCD (GitOps), Kustomize, Istio, and Vault/External Secrets Operator to move services toward container-native, production-parity deployment
- Standardized CI/CD across all services with environment-gated promotions and health-check-triggered rollbacks, cutting deployment cycles by 40%

Software Development Engineer

June 2025 – Oct 2025

InkHub.in / BRMH.ai

Ranchi, Jharkhand

- Built an AI-powered BaaS platform generating full-stack e-commerce apps (RESTful APIs, DB schemas, frontend) from prompts, deployed on AWS (EC2, ECS, Lambda, Step Functions, DynamoDB, Valkey, CloudWatch)
- Re-architected monolithic AWS infra into event-driven, containerized microservices, cutting monthly cloud spend by 76% with zero-downtime migration
- Owned observability and on-call readiness with CloudWatch alarms and structured logging, reducing mean-time-to-detect and accelerating root-cause resolution for production incidents

Software Development Engineer

Nov 2024 – May 2025

EatTogether & Co.

(Remote) Seattle, WA, USA

- Built a sub-100ms dynamic surge pricing engine that personalizes user rates based on real-time availability, spend history, and geolocation
- Refactored a monolith into event-driven microservices with async communication, improving service isolation, independent deployability, and integration with dependent systems
- Collaborated cross-timezone with a distributed US-based team in an Agile/iterative workflow — sprint planning, code review, and incremental rollout of API and pricing changes

TECHNICAL SKILLS

Backend & APIs: FastAPI, Node.js, Express.js, GraphQL, RESTful APIs, OAuth, WebSockets, Celery, SSE, gRPC

Databases & Storage: PostgreSQL, PostGIS, Redis, Milvus, MongoDB, DynamoDB, MinIO

Applied AI & Search: RAG Pipelines, Hybrid Search (Dense+Sparse), RRF Reranking, LangChain, Agentic AI, LLM Integration (Gemini, OpenRouter), MCP, FAISS

Containers & Orchestration: Docker, ECS, Kubernetes, Helm, ArgoCD (GitOps), Kustomize, Istio (Service Mesh), Vault, External Secrets Operator

Cloud (AWS) & IaC: EC2, ECS, ECR, S3, RDS, Lambda, Step Functions, SSM, IAM/OIDC · Terraform, Ansible (Infrastructure as Code)

CI/CD & DevOps: GitHub Actions, automated testing & deployment pipelines, environment-gated promotions, code review practices, unit testing

Observability & Messaging: Grafana, Prometheus, Loki, Tempo, Pyroscope, OpenTelemetry (OTel Collector/Alloy), Alertmanager, Apache Kafka

Languages & Frontend: Python, Go, TypeScript, JavaScript, C++, SQL · Next.js, React.js, Tailwind CSS

Software Engineering: System Design, Microservices & Distributed Systems, Multi-tenant Architecture, CI/CD, SRE, DevSecOps, RBAC, Agile/Scrum

PROJECTS

CaseCloud – AI Legal Case Management

🔗 casecloud.in

FastAPI, PostgreSQL, Milvus, Redis, RAG, Next.js, AWS

- Multi-tenant SaaS with schema-per-tenant isolation (RLS, dynamic search_path) across cases, clients, proceedings, payments, cause lists, and reminders under unified RBAC
- RAG pipeline over private legal corpus — Playwright court portal ingestion, PDF parsing, Milvus vector indexing, hybrid dense+sparse retrieval with RRF reranking and source citations
- Agentic drafting system grounding Gemini over 80M+ Judgements and 6K+ legal templates with placeholder canonicalization, delivering jurisdiction-aware structured drafts from natural language queries

Vedico – Telecom Infra & Realtime Fleet Tracking

🔗 vedicoinfrastructure.com

Go, FastAPI, PostgreSQL, Flutter, Next.js

- Engineered a decoupled Go realtime microservice (goroutine event loop, `sync.RWMutex`-guarded connection maps, channel multiplexing) delivering sub-millisecond "Circle"-scoped WebSocket broadcast of GPS telemetry from field technicians to a NOC dashboard
- Built a Haversine-based spatial distance engine with a 5-minute grace window for offline mobile isolate flushes, plus a theft/proximity anomaly detector combining 50m geofencing, PostgreSQL FOR UPDATE pessimistic locking, and active work-order context to eliminate false positives
- Designed a 4-tier automated SLA escalation state machine (Technician → Supervisor → Cluster Incharge → Manager) with FCM push delivery, transient-failure backoff, and hierarchical RBAC scoping for multi-tenant data isolation

infra-ops – Multi-Environment AWS IaC Platform

🔗 [infra-ops](https://infra-ops.com)

Terraform, Ansible, GitHub Actions, AWS

- Built a centralized Terraform + Ansible IaC repo provisioning dev/staging/prod AWS environments — VPC/subnet/SG networking, EC2 compute, S3, and least-privilege IAM — through reusable, composable Terraform modules
- Implemented GitOps-style CI/CD: PR-triggered `terraform plan`, environment-gated auto-apply to dev with manual approval gates for staging/prod, and a nightly drift-detection job that opens GitHub issues on divergence
- Hardened provisioning with OIDC-based GitHub Actions AWS auth (no long-lived keys), IMDSv2 enforcement, UFW + fail2ban, and encrypted, versioned, DynamoDB-locked Terraform state

lgtm-sidecar – Open-Source Observability Hub

🔗 [lgtm-sidecar](https://lgtm-sidecar.com)

Grafana Alloy, Loki, Tempo, Prometheus, OpenTelemetry, Docker

- Built a production-grade, resource-optimized observability sidecar on the LGTM stack plus Pyroscope and Alertmanager, fronted by Nginx/Certbot TLS, decoupled from application hosts via a lightweight Edge Alloy agent
- Designed OpenTelemetry (Grafana Alloy) pipelines with W3C TraceID correlation across FastAPI and Next.js services, routing `tenant_id/office_id` through span baggage into structured metadata rather than high-cardinality index labels
- Defined SRE-grade operational policy: rate-capped tail sampling, S3-backed Loki/Tempo storage via IAM role, Prometheus/Loki-ruler → Alertmanager alerting (Slack + email), and memory-limited Alloy pipelines to protect host resources

Immigration Copilot – Agentic Form Automation

🔗 [Immigration-Copilot](https://immigration-copilot.com)

Electron, Playwright CDP, LangChain

- LLM agents autonomously navigate and populate immigration forms (I-485, DS-160) via a Playwright CDP browser harness with mid-session decision handling and form state recovery
- LangChain pipeline extracts structured applicant data from noisy documents via schema mapping and field-level conflict resolution — 90% reduction in manual entry

EDUCATION & CERTIFICATIONS

AWS Certified Solutions Architect – Associate (SAA-C03) [\[View\]](#)

June 2026

Master of Science (M.S.) in Computer Science & Data Science (CSDA)

2025-Present

Indian Institute of Technology Patna

Patna, Bihar

Bachelor of Computer Applications

Jul 2025

St. Xavier's College, Ranchi | CGPA: 8.5/10.0

Ranchi, Jharkhand