Hemal Buha

Senior Platform Engineer | AWS | Platform Reliability | AI Infrastructure

Bengaluru, India | buhahemal10@gmail.com | +91-7359943259

https://hemalbuha.pages.dev | https://github.com/buhahemal | https://www.linkedin.com/in/hemalbuha

Professional Summary

Senior Platform Engineer with six years building secure, high-availability multi-tenant SaaS on AWS across backend engineering, cloud infrastructure, and DevOps. Production work spans event-driven backends (Redis Streams), CI/CD modernization, CDN optimization (67% faster asset delivery), API latency reduction (28.7% P95), and a serverless EC2 warm-pool autoscaler with 271+ offline tests. Also ships open-source AI systems with the same engineering discipline: LangGraph multi-agent pipelines, zero-cost embedding stacks, and GitHub Actions automation.

Core Competencies

Platform Engineering | Backend Engineering | Cloud Engineering | DevOps | Distributed Systems | Event-Driven Architecture | AI Infrastructure | High Availability | Performance Engineering | Security Engineering | Infrastructure Automation | Release Engineering | Incident Management | Technical Leadership

Technical Skills

Programming Languages: TypeScript, JavaScript/Node.js, Go, Python, PHP, SQL, Bash
Backend: REST APIs, microservices, API design, webhooks, FastAPI, SDK development, authentication, authorization, SSO
Cloud: AWS Lambda, EC2, ALB, WAF, S3, CloudFront, RDS, SES, SQS, ElastiCache, EventBridge, IAM, CloudWatch, VPC
DevOps: CI/CD, Jenkins, GitHub Actions, Docker, Kubernetes, Argo CD, release engineering, deployment automation, Lighthouse CI
Infrastructure: CloudFormation, Linux, autoscaling, zero-downtime deployments, disaster recovery, cloud cost optimization
Databases: MySQL, MariaDB, MongoDB, Redis, Amazon RDS, Supabase, caching, database performance tuning
Distributed Systems: Redis Streams, asynchronous workers, queues, event-driven architecture, WebSockets, distributed messaging
AI: LangGraph, LLM orchestration, multi-agent systems, sentence-transformers, embeddings, vector search, semantic search, AI assistants
Frontend: React, Next.js, Vite, TypeScript
Security: OWASP, JWT, OAuth, secure webhooks, VAPT remediation, GDPR, ISO 27001, input validation, access control
Architecture: Multi-tenant SaaS, white-label SaaS, scalable backend architecture, high availability, API integration patterns
Monitoring: CloudWatch, centralized alerting, logging, production monitoring, incident response
Operating Systems: Linux
Networking: AWS VPC, DNS, HTTP/HTTPS, TLS

Professional Experience

Edmingle | *Senior Platform Engineer* | Bengaluru, India — Jul 2021 – Present

- Architected cloud-native backend capabilities for a multi-tenant, white-label SaaS platform using Node.js, TypeScript, AWS, Redis, and MySQL, balancing customer-specific integrations with reusable platform patterns.
- Designed and deployed a serverless EC2 warm-pool autoscaler (Node.js 24 / Lambda) for an API fleet behind ALB: desired-state reconciliation, CPU scaling ladder, self-healing, WAF IP-set sync, and 271+ offline tests.
- Engineered a Redis Streams concurrency framework with asynchronous workers, retry handling, and alerting for three burst-write domains—signup, enrollment, and social workloads—creating a reusable event-driven processing model.
- Modernized release engineering by rebuilding GitHub Actions CI/CD pipelines and staging major database upgrades across private and shared environments without service interruption; the pipeline changes saved roughly 10 minutes per build.
- Directed the Bunny-to-CloudFront asset migration with S3 and AWS services, reducing average asset response time by 67% while coordinating delivery dependencies across engineering teams.
- Optimized high-traffic APIs through read-path simplification, Redis caching, and RDS index cleanup, cutting targeted API P95 latency by 28.7% and the highest-impact endpoint by 54%.
- Strengthened platform security through VAPT remediation, IDOR and input-validation fixes, secure webhook idempotency, and authentication/authorization hardening; contributed to ISO 27001 and GDPR readiness reaching 100%.
- Hardened payment and communications reliability across four payment gateways—including Razorpay, Stripe, and PayU—and AWS SES by implementing duplicate-webhook safeguards, settlement alerting, private email infrastructure, and a permanent Microsoft/Outlook deliverability fix.
- Guided architecture reviews, production incident response, delivery planning, and code-quality practices for a three-person infrastructure/backend team; mentored two engineers and enabled support self-service through secure API-key tooling.

Rovae Inc. | *Software Engineer* | India — Jul 2020 – Jul 2021

- Developed Laravel backend services and REST APIs backed by MySQL for inventory-management and point-of-sale applications, connecting multiple third-party systems to dependable business workflows.
- Refined application maintainability and deployment workflows through focused backend enhancements, making the codebase easier to operate and extend.

Selected Engineering Projects

EC2 Warm Pool Autoscaler Production | Platform Engineering

- Built AWS Lambda autoscaler (Node.js 24) for a fixed EC2 warm pool behind ALB: desired-state reconciliation, CPU scaling ladder, self-healing, WAF IP-set sync, and 271+ offline tests with dry-run rollout.
- Replaced manual cron-based scaling and unreliable ASG behavior; EventBridge schedules and 3-minute converge loop with idempotent start/stop/register/drain operations.

Haat - Multi-Agent LangGraph Research Desk Open Source

- Architected 14+ LLM agent pipeline (analysts, bull/bear debate, trader, risk stress-test, portfolio manager) producing structured ratings and machine-readable trade tickets.
- Grounded LLM stages with deterministic 6D quant engine and ATR ticket math; FastAPI/SSE web desk, Typer CLI, checkpoint/resume; 565 pytest cases (github.com/buhahemal/haat).

Parkhu AI Open Source | Market Intelligence

- Pre-open institutional research pipeline for Indian equities: 13 collectors, 9 derived steps, gated swing brief with suggestion ledger, GitHub Actions, and 366-symbol universe (github.com/buhahemal/parkhu-ai).

AI Job Hunter Open Source | Job Discovery

- Zero-cost job discovery monorepo: 10+ ATS scanners, local sentence-transformers embeddings, Supabase storage, and React dashboard on GitHub Pages (github.com/buhahemal/ai-job-hunting).

MyLumo Website Production | mylumo.in

- Live Next.js static-export marketing platform: 18 routes, 38 MDX blog posts, build-time SEO pipeline, GitHub Actions deploy to S3 + CloudFront with Lighthouse CI (mylumo.in).

ZyncSpace - AI-Powered Collaboration Platform Personal Engineering Project

- Designed a scalable backend for real-time collaboration using Node.js, AWS, Redis, Redis Streams, and WebSockets, separating distributed messaging from interactive application flows.
- Integrated semantic search, embeddings, vector search, and an AI assistant to provide contextual collaboration features while designing for high availability and operational reliability.

Selected Engineering Achievements

- Delivered a staged MariaDB major-version upgrade with zero data loss and less than five minutes of downtime in 2024; subsequently orchestrated zero-downtime fleet upgrades across private and shared environments.
- Reduced live API response time by 70% by moving daily-statistics processing into background jobs, and cut a batch-progress API from timeouts to under one second.
- Implemented AWS SES-based email delivery and custom-domain routing, reducing email delivery costs by approximately 40%; added abuse mitigation and bounce monitoring to protect sender reputation.
- Executed disaster-recovery activities with an RTO under 15 minutes and RPO under 10 minutes, reinforcing platform resilience and operational readiness.
- Automated new-client onboarding and Vimeo domain whitelisting, reducing manual operational work by 80% while strengthening reliability of video access.
- Introduced Go workers with Amazon SQS and queue-based processing for high-concurrency transfers and requests, raising throughput and isolating peak-load work from synchronous APIs.
- Hardened payment APIs with authentication, input validation, and rate limiting, and closed high- and critical-severity VAPT findings within SLA.

Education

Master of Computer Applications (Cloud Technology) | *Ajeenkya DY Patil University* | India CGPA: 9.8 / 10
Gold Medalist