BALRAM SINGH

Solutions Architect — Cloud & AI Platforms

Gurugram, Delhi NCR, India | +91 79831 65052 | balram@integrably.com | linkedin.com/in/balramsingh-tech | balram.integrably.com

PROFESSIONAL SUMMARY

Solutions Architect with 6+ years designing and delivering cloud-native and AI platforms end-to-end — from technical discovery and stakeholder workshops to hands-on build and production deployment. Bridge business challenges and engineering capability, architecting agentic AI, RAG, and event-driven systems on AWS with cloud-portable designs spanning GCP and Azure. Delivered an ISO 27001 / SOC 2-aligned, multi-tenant fintech AI platform processing $50M+ in invoices and API infrastructure serving 100K+ daily production calls (load-tested to 10K+ RPS burst), while leading a cross-functional team of 8 engineers.

CORE SKILLS

Cloud & Infrastructure: Google Cloud Platform (GCP), AWS (Lambda, S3, EC2, DynamoDB, API Gateway, Textract, SageMaker), Azure, Docker, Kubernetes, Serverless, Event-Driven Microservices

AI / ML: Agentic AI Workflows, RAG, LLMs (self-hosted & API), LangChain, LangGraph, MCP, CrewAI, Vector Databases, Embeddings, vLLM/TGI, LLM Evals, OCR

Languages: Python, Java, JavaScript/TypeScript, SQL

Backend & Data: Java 21 (Virtual Threads), Spring Boot, FastAPI, Apache Kafka, REST APIs, Microservices, Oracle, MySQL, PostgreSQL, MongoDB, Redis, Caffeine

Architecture & Delivery: Solution Architecture, Multi-Tenant SaaS (Schema-per-Tenant), Technical Pre-Sales / PoV Delivery, Stakeholder Management, Terraform, CI/CD (Jenkins, GitHub Actions), Security (ISO 27001, SOC 2, RBAC), Agile/Scrum

PROFESSIONAL EXPERIENCE

Lead Solutions Architect & Technical Manager | TanServ — Jan 2025 – Present

Gurugram, India · Enterprise Fintech SaaS (AI-driven Accounts Receivable)

- Architected a zero-downtime, multi-tenant AR SaaS on AWS — Java 21 (virtual threads)/Spring Boot with an event-driven AI pipeline (S3 → Lambda → OCR → inference → DynamoDB → API Gateway), designed cloud-portable (1:1 GCP mapping).
- Reduced manual operational workload by 60% by designing agentic AI workflows (LangGraph, MCP) and a RAG system over a custom vector-database layer built on existing relational data.
- Cut AI infrastructure cost ~40% at target p95 latency by offloading OCR to Amazon Textract — freeing GPU capacity for self-hosted LLM inference via vLLM/TGI — after a managed-vs-self-hosted TCO evaluation.
- Replaced OLTP polling with Apache Kafka event streams and a two-level cache (Redis + Caffeine) for real-time dashboards; 99.95% uptime via MySQL regional HA (synchronous standby + read replicas) and daily zero-downtime releases.
- Spearheaded ISO 27001 / SOC 2 readiness: schema-per-tenant database isolation with per-client encryption keys, automated audit versioning across 150+ domain entities, RBAC, and encryption in transit and at rest.
- Lead and mentor a cross-functional team of 8 engineers (5 backend Java/Spring Boot, 3 frontend React/Angular), partnering directly with the founder and product managers to translate business objectives into scalable technical requirements.

Senior Software Engineer | Safexpress — Jun 2023 – Jan 2025

Gurugram, India · Mission-critical logistics & supply-chain systems

- Reduced API response times by 85% (800ms → 120ms) by optimizing complex Oracle/SQL queries across high-scale supply-chain services.
- Scaled backend logistics services to 50K+ daily transactions at 99.9% uptime by deploying microservices on AWS (EC2/S3) with automated Jenkins CI/CD, expanding automated test coverage from 40% to 85%.
- Enabled real-time shipment tracking by building a WebSockets + Redis notification system and integrating third-party logistics-partner APIs.

Full-Stack Developer (Contract) | Mastercard — Jan 2023 – Jun 2023

- Delivered 12+ enterprise features for global payment-processing reporting with zero production incidents, and improved dashboard load time by 45%.

Software Engineer (Java) | Techginia Global — Nov 2020 – Dec 2022

- Architected and deployed secure REST APIs on AWS serving 20K+ daily active users at 99.7% uptime, and cut deployment time from 2 hours to 15 minutes.

SELECTED PROJECTS

- **Inebura (Fintech AI Platform) —** Architected an AI-powered accounts-receivable platform processing $50M+ in invoices at 95%+ extraction accuracy, cutting manual work 70% and yielding $500K+ in annual savings.
- **Enterprise API Gateway —** Engineered a centralized gateway with auth, distributed rate-limiting (Redis) and routing — 100K+ daily production requests, load-tested to 10K+ RPS burst at <50ms p95 latency.

EDUCATION

Bachelor of Computer Applications (Hons.) | Institute of Technology and Management, Gwalior | 2017 – 2020