

BALRAM SINGH

Solutions Architect — Cloud & AI Platforms

Gurugram, Delhi NCR, India | +91 79831 65052 | balramthewarrior@gmail.com | [linkedin.com/in/balramsingh-tech](https://www.linkedin.com/in/balramsingh-tech) | balrams.in

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 across GCP, AWS, and Azure. Delivered a fintech AI platform processing \$50M+ in invoices and API infrastructure sustaining 100K+ requests/sec, 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, CrewAI, Vector Databases, Embeddings, vLLM/TGI, OCR

Languages: Python, Java, JavaScript/TypeScript, SQL

Backend & Data: Spring Boot, FastAPI, REST APIs, Microservices, Oracle, MySQL, PostgreSQL, MongoDB, Redis

Architecture & Delivery: Solution Architecture, Technical Pre-Sales / PoV Delivery, Stakeholder Management, 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 an end-to-end, event-driven AI platform on cloud (S3 → Lambda → OCR → model inference → DynamoDB → API Gateway), transforming the product from reactive software into an autonomous, predictive AR ecosystem.
- Reduced manual operational workload by 60% by designing agentic AI workflows and a LangChain + RAG system over a custom vector-database layer built on existing relational data.
- Cut infrastructure cost ~40% at target latency by evaluating managed vs. self-managed options and deploying a self-hosted multiple LLM on GPU (vLLM/TGI), reserving GPU compute for inference via Amazon Textract OCR etc.
- Sustained 99.95% platform stability during rapid scaling by optimizing deployment pipelines and automated testing guardrails, moving releases from weekly to daily.
- Brought the platform into ISO 27001 / SOC 2 alignment by implementing RBAC, encryption in transit and at rest, audit trails, and access management.
- 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 pipelines.
- 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 0 by implementing Load Balancers – shifting traffics.

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 management gateway sustaining 100K+ requests/sec at <50ms latency.
- ComixFlip** — Scaled a digital platform with AI-driven recommendations to 10K+ users across a 50K+ asset library.

EDUCATION

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