TITAS BISWAS

Technical Architect — Backend & AI Systems

+91 9903423090 | titas.11g@gmail.com | titasbiswas.com | linkedin.com/in/titasbiswas-94b79066 | github.com/phoenixtb

Bangalore, India

SUMMARY

Hands-on technical architect with 16 years of experience on designing and developing high-availability, event-driven systems for global companies (Tesco, VMware). Now focused on applying deep expertise in distributed systems to architect and integrate production-grade AI and Machine Learning solutions into the enterprise. Recently completed an Executive PGP in AI & ML from IIIT Bangalore and developed a portfolio of AI projects demonstrating practical skills in RAG, LLM integration, and edge computing. Seeking a challenging role to architect and lead the integration of AI/ML capabilities into enterprise applications, leveraging a strong foundation in cloud-native technologies, real-time data streaming, and edge computing.

TECHNICAL SKILLS

Core: Java, Scala, Spring Boot, Spring WebFlux, Apache Kafka, Kubernetes, Docker, K3s, AWS, Terraform, OpenTelemetry, Prometheus

Databases: Couchbase, PostgreSQL, MongoDB, Amazon DynamoDB, Apache Cassandra, ElasticSearch, Redis

Architecture: Microservices, Event-Driven Architecture, Edge Computing, Distributed Systems, System Design, CQRS, CDC, Stream Processing

AI/ML: RAG Systems, Haystack, Qdrant, LiteLLM, LangGraph, On-device Inference, Vector Databases

API & Integration: REST, GraphQL, Server-Sent Events, , ONVIF, RTSP

AI, PERSONAL PROJECTS & PUBLICATIONS

DocIntel: Enterprise RAG & Knowledge Engine | https://github.com/phoenixtb/docintel

- A production-grade, multi-tenant RAG platform with full tenant data isolation (per-tenant vector collections, PostgreSQL RLS, object storage buckets) and declarative RBAC via OPA; RAG pipeline features hybrid retrieval (dense + BM25/RRF), cross-encoder reranking, semantic response caching, zero-shot domain routing, and ACL-level access control per role — traced end-to-end via Langfuse.
- **Technology:** Kotlin/Spring Boot · Python/FastAPI/Haystack · Qdrant · PostgreSQL · MinIO · ClickHouse · Redis · OPA · Authentik · SvelteKit · Docker

Edumate Lite: On-Device AI Tutoring | https://github.com/phoenixtb/edumate-lite-android

- Fully offline, privacy-first AI education assistant running multi-model inference locally on Android via dual engine support (LiteRT + llama.cpp/GGUF); RAG pipeline ingests PDFs and scanned documents, embeds into a local vector store, and retrieves semantically — zero network calls. Agentic study mode runs multi-step ReAct loops with memory-pressure-aware iteration gating, orchestrating tools for search, summarization, and worksheet generation against the same local inference stack.
- **Technology**: Kotlin/Compose · Koin · Room · LiteRT · llama.cpp · JetBrains Koog · CameraX/ML Kit OCR · Android

Portfolio & AI Chat | titasbiswas.com

- Personal portfolio with a RAG-powered AI chat — SvelteKit deployed on Cloudflare Workers, custom ingestion pipeline chunks pub site content into Vectorize for semantic retrieval, with incremental re-embedding and IP rate limiting

AI: Through an Architect's Lens | Blog Medium

- Publishing tutorial series on Medium bridging distributed systems with AI/ML.

EXPERIENCE

Technical Architect (SDE 3) — Tesco, Bangalore

Feb 2023 – Present

Tech: Java, Scala, Spring WebFlux, Apache Kafka, Couchbase, Kubernetes, K3s, Prometheus,

Architecture lead for two products under the Safe & Secure Stores umbrella.

Video Platform (current) — ~1,500 stores, ~45,000 cameras

- Leading architecture and development for the video platform supporting all express stores with planned expansion to large-format stores. Platform runs on edge K3s clusters with selective cloud export.
- Driving the Total Loss Recovery (TLR) programme — linking video evidence with loss events for recovery workflows.
- Proposed and leading edge camera analytics initiative — designing ONVIF Profile T/S event collection from Hanwha cameras on K3s edge clusters with bandwidth-constrained cloud export (1 Mbps). Engaged directly with Hanwha for technical collaboration.

Incident Prevention Platform (Feb 2023 – Nov 2025)

- Designed and built real-time alerting system (Server-Sent Events + Spring WebFlux + Apache Kafka) delivering notifications for ~1.5 million ANPR events/day and ~3,000 fraudulent transactions/day to ~200 security operators and store managers.
- Architected and lead the development of incident reporting and evidence-linking system for store-level loss prevention and prosecution.

Staff Engineer — **VMware, Bangalore** Feb 2022 – Jan 2023

Tech: Java, Scala, Apache Spark, Spring Boot, Amazon DynamoDB, Apache Kafka, AWS

- Designed and led CloudHealth — multi-cloud SaaS platform with event-driven microservices processing ~100 TB/day for cloud cost management and optimisation.

Staff Software Engineer — **Bazaarvoice, Bangalore** Jun 2020 – Feb 2022

Tech: Java, Scala, Finatra, ElasticSearch, Apache Cassandra, Amazon DynamoDB, AWS ECS, AWS Kinesis, Docker

- Led Curalate content platform team through acquisition transition; system handled 3 billion+ API requests during 36-hour Black Friday peaks.
- Designed and developed media processing platform handling ~35,000 images/day, replacing legacy system.

Senior Software Engineer — **Altimetrik, Bangalore** Apr 2019 – May 2020

- Built high-performance async microservices with distributed tracing (Spring Sleuth, Zipkin).

Advisory Systems Analyst — **IBM, Kolkata** Jun 2016 – Mar 2019

- Designed enterprise applications for power utilities and airline hospitality domains; led small delivery team.

Earlier: Cognizant (2014–2016), Siemens (2011–2014), Infosys (2008–2011) — Enterprise Java development.

EDUCATION

Executive PGP in AI & Machine Learning — IIIT Bangalore 2024–2025 (Awaiting Certificate)

Bachelor of Technology — Jadavpur University 2004–2008

Portfolio: titasbiswas.com | GitHub: github.com/phoenixtb | Medium: medium.com/@titas.11g | Blog: pub.titasbiswas.com | GitLab: gitlab.com/titas.biswas