VIKAS KUMAR SINGH

Principal AI Engineer & Systems Architect

Bangalore, India • singhvks@outlook.in • linkedin.com/in/singhvks • github.com/singhvks

PROFESSIONAL SUMMARY

AI systems architect owns production-grade multi-agent platforms and LLM infrastructure from design through operation. Proven at architecting for 1000+ concurrent users, billion-scale data pipelines, and sub-100ms inference latency under cost constraints. Core expertise: async backend design (FastAPI, connection pooling, fault tolerance), LLM orchestration (routing, memory management, token budgeting), and distributed systems (Spark, Databricks, cloud infrastructure). Track record: prevented $2B supply chain waste, automated $100K operational overhead, reduced compute costs 65% without quality loss. Operates autonomously: architect, design, build, ship, scale - no delegation.

CORE ARCHITECTURE

- **LLM Systems at Scale:** Multi-agent architectures (LangGraph), agentic orchestration & routing, RAG with retrieval optimization, prompt caching & token budgeting, async inference with graceful degradation, latency SLAs, output validation & guardrails
- **Backend Infrastructure:** Async FastAPI (concurrent request handling, batching, connection pooling, caching), Python async patterns & event loops, Docker/K8s, AWS (SageMaker, Redshift, Glue), GCP (Vertex AI, Bigquery), Azure; sub-100ms latency targets, SSO Authentication
- **Distributed Data Platforms:** Databricks Lakehouse (Unity Catalog governance), PySpark (billion-row pipelines, dynamic partitioning), Delta Lake optimization, data modeling for OLAP at scale, cost-per-query optimization
- **Regulated Systems (HIPAA/GDPR/Solvency II):** End-to-end compliance architecture: PII detection & masking, field encryption, audit logging, RBAC, data retention policies, automated compliance validation

PROFESSIONAL EXPERIENCE

Data Science Manager • Tredence Analytics, Bangalore (Nov 2024 – Present)

- **Owned multi-modal analytics platform (FastAPI + LangGraph + Databricks)**: architected for 1000+ concurrent users at <100ms latency; replaced 40 FTE manual reporting with agentic workflow orchestration (60 reports/week automated). Economic impact: $100K/year operational overhead elimination.
- **Designed and implemented a comprehensive cost-governance layer** (token budgeting, semantic caching, dynamic LLM routing), cutting inference costs by 65% ($8 to $2.50/request) while upholding high output quality; scaled architecture to support 10K concurrent users.
- **Owned 1TB+ geospatial Lakehouse:** migrated from schema-on-read Hive to star-schema Delta + Unity Catalog; designed dynamic partitioning strategy for 200M geospatial features. Query latency: 45sec→<2sec (22x improvement). Reduction in site-selection approval cycles: 3 weeks→5 days (70% acceleration). Direct business impact: $50M+ CAPEX allocation decisions informed monthly.
- **Owned integration layer:** built seamless translation between AI backend (async FastAPI, Kafka message queue) and frontend systems (React, Tableau, Excel); engineered request deduplication to handle enterprise idempotency requirements.

- Led **cloud migration & cost optimization**: decommissioned legacy on-prem ETL, migrated to AWS Glue + Databricks. Designed compute governance framework (auto-scaling policies, workload isolation, cost allocation by business unit). Result: 30% cloud cost reduction, eliminated $1M/year system maintenance overhead. Owned this autonomously across 6-month implementation.
- **Served as SME and implementation architect for GenAI/Analytics solutions,** driving platform readiness and providing hands-on guidance for team enablement, phased solution rollout, and scaling roadmaps.

Business Technology Solutions Consultant, AI Systems • ZS Associates, Pune (Apr 2021 – May 2024)

- Architected **demand forecasting engine** for $10B+ pharma: chose Bayesian MCMC over standard econometrics (faster convergence on irregular patterns, native uncertainty quantification). Owned end-to-end: model R&D, Spark feature pipeline (500M+ SKU-day pairs), inference orchestration, MLOps. Economic outcome: $2B supply chain waste prevention over 5 years.
- **Implemented Spark-based recommendation engine**: engineered automated pipelines to deliver intelligent next-best-action insights for MSLs, optimizing KOL engagement and touchpoint frequency.
- **Scaled AI delivery organization**: architected MLOps platform (MLflow versioning, automated model validation, A/B testing framework, audit logging) enabling 28 distributed engineers to ship 6 concurrent enterprise projects without workflow bottlenecks. Built a compliance automation layer ensuring HIPAA/GDPR validation on every model commit.

Data Engineer • Collabera Technologies, Pune *Sep 2020 – Apr 2021*

- **Architected HIPAA-compliant Real-World Data (RWD) platform**: owned PII detection, field-level encryption, audit logging. Processed 500M+ healthcare records/day; maintained <10MB per-patient data footprint while preserving analytics capability. This application ingested data from Komodo for HCP scoring.
- **Eliminated BI tool licensing costs ($2M+ contract cycle)**: engineered custom Python analytics layer replacing vendor dependency. Designed query optimization for 100M+ patient records with sub-second response on aggregate queries.

Engineer • L&T Infotech, Pune *Sep 2016 – May 2020*

- **Owned ML-driven insurance risk platform:** architected API-integrated model serving (sub-50ms inference SLA), built multi-armed bandit A/B testing framework for 10+ concurrent models, implemented feature store for 500+ features with drift detection.
- **Engineered GDPR/Solvency II compliant ETL:** owned Spark pipelines handling regulated financial data. Designed audit-proof architecture: immutable transaction logs, cryptographic data lineage, automated compliance validation on every pipeline run.

CERTIFICATIONS & EDUCATION

- **Databricks Certified Generative AI Engineer** - 2025
- **Generative AI Solutions Architect** - 2025
- **Certified NLP Developer** - 2018
- **Certified Python Developer** - 2017
- **Master of Science, Applied Data Science** - WorldQuant University
- **B.Tech, Electronics Engineering** - BVDU College of Engineering, Pune (2016)