VIKAS KUMAR SINGH

AI Architect | Generative AI Platform
singhvks@outlook.in | linkedin.com/in/singhvks | github.com/singhvks |
https://singhvks.github.io/Vikas-CV/

PROFESSIONAL SUMMARY

Principal AI Architect with 10 years of building production-grade AI and data platforms inside large regulated enterprises including pharma, retail, healthcare, and financial services. Delivered outcomes include ~$2B projected pharma supply-chain savings, $50M+ monthly CAPEX decisions on real-time data, and 65% LLM inference cost reduction. Seeking Solutions Architect role to drive strategic GenAI adoption and lead high-impact engineering initiatives for top-tier clients.

BUSINESS IMPACT HIGHLIGHTS

- $50M+ monthly CAPEX decisions on real-time data following 22x geospatial query latency improvement and 70% acceleration in site-selection approval cycle.
- 65% LLM inference cost reduction on multi-agent GenAI analytics platform serving 1,000+ concurrent enterprise users.
- 40 FTE manual reporting automated via agentic workflow orchestration; $1M+/year cloud maintenance overhead eliminated.
- ~$2B projected drug supply waste avoidable over 10-year horizon using Bayesian MCMC clinical trial forecasting.

CLOUD ARCHITECTURE & GENAI EXPERTISE

GenAI & LLM Systems: Multi-agent orchestration, RAG architecture, LLM-as-Judge, token budgeting, semantic caching, dynamic model routing, prompt engineering.
Data Platforms: Databricks Lakehouse, PySpark, AWS Glue & S3, CloudWatch MLOps, DuckDB, OLAP design, geospatial feature management.
ML & Statistics: Bayesian MCMC, XGBoost, SHAP explainability, hierarchical priors, MICE imputation, Poisson process simulation.
MLOps & Engineering: MLflow, Docker, Kubernetes, async FastAPI, CI/CD pipelines, HIPAA / GDPR compliance automation.
Languages: Python, SQL, PySpark

PROFESSIONAL EXPERIENCE

Tredence Analytics | Bangalore, India | Nov 2024 – Present | *Data Science Manager & Principal AI Architect*

- Led architecture and design of a production GenAI analytics platform, scaling it to 1,000+ concurrent enterprise users with <100ms latency, automating 60 reports/week and eliminating $100K/year operational overhead.
- Implemented LLM cost-governance layer (token budgeting, caching), cutting inference costs by 65% while maintaining quality.
- Directed and executed a 1TB+ Geospatial Lakehouse migration, improving query latency by 22x (45s -> 2s) and accelerating site-selection approval from 3 weeks to 5 days.
- Decommissioned legacy ETL and migrated to AWS Glue + Databricks, resulting in 30% cloud cost reduction and $1M+/year maintenance savings.

ZS Associates | Pune, India | Apr 2021 – May 2024 | *Business Technology Solutions Consultant, AI Systems*

- **Safety Stock Forecasting for Phase 2/3 Trial:** Developed Bayesian MCMC engines to predict site enrollment and drug supply needs for high-stakes Oncology launches, ensuring optimal stock preparation and reducing wastage.
- **Launch Engagement Optimization:** Implemented Spark and ML based recommendation engines to optimize MSL touchpoints during pre-launch phases, tracking engagement metrics to ensure multi-territory launch readiness.
- **Clinical Trial Prioritization:** Engineered NLP-based scoring architectures to streamline budget distribution and trial request ranking. By integrating disparate global datasets, I accelerated the preparation timelines for upcoming drug launches and enabled data-driven prioritization of trial submissions.
- **Omnichannel Analytics Reporting Platform:** Architected automated AWS data pipelines and optimized data models to track critical drug launch preparation and success metrics. Transformed raw data into actionable commercial insights via Tableau, accelerating executive decision-making during high-stakes drug launch cycles.

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

- **Real-World Data (RWD) Diagnostic Tool:** implemented integration of EHR and claims data into a centralized dashboard to automate feasibility assessments for acquired datasets. This initiative streamlined data richness evaluations by 60% and achieved $1M+ in annual savings by transitioning from Tableau to a custom Python Plotly solution featuring automated quality checks and email alerts.

L&T Infotech | Pune, India | Sep 2016 – May 2020 | *Engineer - AI & ML Systems*

- Developed a recommendation REST API for breakdown cover to provide tailored policy suggestions using user profiles and historical data, successfully minimizing user drop-off by 27%.
- Engineered ETL for Solvency compliance and reporting for transactions on the policy admin system.

EDUCATION

- **M.Sc. Applied Data Science** | WorldQuant University | 2021
- **B.Tech Electronics Engineering** | BVDU College of Engineering, Pune | 2016
- **Diploma Network Security** | BVDU College of Engineering, Pune | 2015

CERTIFICATIONS

- Databricks Certified Generative AI Engineer (2025)
- Generative AI Solutions Architect (2025)
- Certified NLP Developer (2018)
- Certified Python Developer (2017)
- AI/ML for Geodata Analysis - ISRO