Subhodeep Sen

Generative AI Architect (Applied AI, GenAI, MLOps)

+91 9836647441 | subhodeep.sn@gmail.com | Bangalore - 560075
https://linkedin.com/in/subhodeep-sen

Professional Summary

Strategic and innovation-driven technology leader with experience in building and scaling **data platforms, AI/ML-enabled products**, and cloud-native microservices for enterprise advisory and consumer-scale use cases. Specializing in Applied AI/ML, GenAI (RAG/agentic), and enterprise-grade governance, security, and reliability. Currently leading architecture and delivery for Deloitte Advisory (M&A/Valuation) platforms—shipping production services that embed intelligence into workflows and enabling responsible GenAI adoption through reusable guardrails, traceability, and policy enforcement patterns.Expertise in **Data Engineering Pipelines, Streaming Data, Spark, Agentic Architecture, MLOPs, LLMOps, Generative AI System Design, Google Cloud, AWS and Microservices Technology stack**.Strong hands-on background in **Python, NLP, Deep Learning, SQL and Distributed Systems and Data Streaming Architectures** (Kafka, Spark, Airflow), and partnering closely with Product and business leaders to translate ambiguous problems into well-defined data/ML solutions, measurable outcomes, and execution-ready roadmaps in regulated, risk-adjacent environments.

Achievments

- Led diverse teams and delivered **complex large-scale products** on major public cloud environments.
- Led end-to-end design and rollout of **AI-enabled digital products** (valuation and advisory platforms) used across multiple member teams, aligning **product outcomes, architecture, and delivery execution**
- Built **GenAI guardrails** for **RAG and agentic architectures** (LangChain/LangGraph + vector search patterns), focusing on **safety, traceability, and responsible AI controls** for enterprise use
- Deep expertise in **cloud-native data engineering** (Spark/Flink/Airflow/Kafka) and **streaming-first architectures**, enabling reliable ingestion, processing, and analytics at scale
- Proven ability to translate business problems into technical designs, rapid prototypes, and production releases through tight collaboration with **product, UX, and engineering**
- Strong background in **microservices, async Python**, real-time communication APIs, and distributed caching/search for low-latency, high-traffic systems
- Ownership of **platform reliability and delivery automation** using **Docker/Kubernetes** and CI/CD/DevOps patterns; emphasis on observability, fault tolerance, and operational excellence
- Certifications: **AWS Certified Solutions Architect**; **Microsoft Certified Azure Data Engineer Associate**

Core Competencies

- **Applied AI / GenAI:** RAG pipelines, agentic workflows, **guardrails frameworks**, vector databases, LangChain, LangGraph, Google ADK, prompt orchestration patterns, LLMOps concepts, LLM Fine Tuning of Mistral with Unsloth platform on valuation data ,evaluation, safety red-teaming, RLHF/DPO,Model governance, AI Ethical compliance.

- **NLP:** TF-IDF, KNN, feature engineering foundations, evaluation-oriented prototyping, unstructured text processing, document parsing, extraction, entity linking, contract analysis, scikit-learn, nltk, gensim
- **Core Machine Learning:** Deep Learning Frameworks like TensorFlow 2.x, TFRS (TensorFlow Recommenders), Keras, PyTorch. Knowledge graph search, Hybrid search and Graph Neural Networks,**Autoencoders**, Anomaly Detection.
- **RecSys Architectures**: Two-Tower Retrieval, Deep & Cross Networks (DCN-V2), Wide & Deep Learning, Sequential Models (RNN/Transformers), Multi-Task Learning (MTL), **Transformers.**
- **Optimization:** Hyperparameter Tuning (Keras Tuner), Post-Training Quantization, Gradient Clipping, Embedding, Regularization.
- **MLOps & Engineering:** Pipeline Orchestration , TFX (TensorFlow Extended), Kubeflow Pipelines, Airflow, MLflow.TensorFlow Serving, FastAPI, Docker, Kubernetes (K8s), Helm.
- **Monitoring & Validation:** TFDV (Data Validation), TFMA (Model Analysis), Prometheus, Grafana, Model Drift Detection.
- **Data Engineering:** Spark, Flink, Airflow,MapReduce, Kafka, Kafka Streams,batch + streaming pipeline design, data ingestion patterns, pandas, Azure ML and Databricks
- **Backend & APIs:** Python 3, Asyncio, Sanic, Flask, REST, SOAP, WebSockets, microservices, asynchronous programming, service discovery patterns
- **Cloud Platforms:** Google Cloud Platform (Vertex AI, GKE), AWS (SageMaker, EKS).
- **Datastores:** Postgres, MySQL, MS SQL Server; MongoDB, Cassandra, HBase, Redis, Memcached, Big Query , Databricks , graph database concepts
- **Leadership:** Engineering management, hiring/mentoring, agile delivery, stakeholder management, cross-functional execution, roadmap planning, architecture governance, GTM support

Work Experience

Deloitte Bangalore, Nov 2018 - Present
AI/ML Lead Architect (Applied AI, GenAI, MLOps)

- **Designing** Highly Reliable Scalable Fault Tolerant Systems and Deployment Pipelines for critical valuation and financial modeling products in **Risk and Financial Advisory products group**
- Recruit, hire, mentor and manage teams of software engineers. Manage the agile development process and methodology to deliver tech requirements on time and with a **high degree of precision and quality**.
- **Developed Guardrails Service (**GRS) - guardrail framework for RAG and Agentic Architectures
- **Spearheaded** an Engineering Team to launch cloud-based **Valuation and Modeling products** to carry out automated Valuations for organizations going for merger and acquisition. The product scales across **several member teams** all through Deloitte currently.
- **Architected** an **Enterprise-Scale Hybrid Discovery Engine** using two-stage recommendation pipeline (Retrieval and Ranking) using TensorFlow Recommenders (TFRS) to provide personalized content to **100K+ active users**.
- **Developed** a Two-Tower Retrieval model to generate high-quality candidate sets from a library of **1M+ items**, leveraging user-item interaction embeddings and metadata.
- **Architected** a Deep & Cross Network (DCN-V2) for the ranking stage, improving the model's ability to learn complex, non-linear feature interactions without manual feature engineering.
- **Optimized** inference latency by **85% (from 200ms to 30ms)** by integrating ScaNN (Scalable Nearest Neighbors) for efficient vector similarity search in production.
- **Desgined** a robust evaluation framework utilizing Mean Reciprocal Rank (MRR) and Recall@K metrics, leading to a **14% uplift in user retention** over a 6-month period.

- **Reduced training time by 40%** by implementing custom tf.data input pipelines and utilizing mixed-precision training on NVIDIA A100 GPUs.
- **Built an end-to-end MLOps ecosystem** using TensorFlow Extended (TFX) and Kubeflow, automating the transition from experimental code to production-ready microservices.
- **Integrated TensorFlow Data Validation (TFDV)** into the ingestion layer to automatically detect schema drift and training-serving skew, preventing degraded models from reaching production.
- **Designed a Continuous Training (CT)** trigger that monitors model performance in real-time and initiates automated retraining workflows when precision drops below a defined threshold.
- **Leveraged TensorFlow Model Analysis (TFMA)** to perform "Slicing Analysis," ensuring model fairness and consistent performance across different user demographics (e.g., age groups, regions).
- **Orchestrated a "Champion-Challenger" deployment strategy**, allowing for safe A/B testing of new model versions with zero downtime for the end user.
- **Reduced the model deployment lifecycle by 80%** and eliminated **90% of manual data-cleaning** tasks through automated validation.
- **Project - High-Throughput Real-Time Feature Serving for Valuation Products : Architected** a real-time inference engine using TensorFlow Serving on Kubernetes (GKE), successfully managing peak traffic of **5,000+ requests per second** (RPS).
- **Developed** a low-latency Feature Store integration using Redis, enabling the model to fetch and inject live user session data (e.g., last 5 clicks, current balance) into the inference request in under 10ms.
- **Optimized model deployment** via Post-Training Quantization (PTQ), reducing the model's memory footprint by **4x** and enabling faster cold starts on autoscaled Kubernetes pods.
- **Implemented an API Gateway** using FastAPI to handle request validation, logging, and asynchronous data enrichment before passing payloads to the TensorFlow Serving cluster.
- **Ensured 99.9% system availability** by configuring Horizontal Pod Autoscaling (HPA) based on custom Prometheus metrics (CPU/Request Latency)
- **Achieved a 15%** increase in conversion for time-sensitive financial offers by delivering personalized recommendations in under **100ms.**
- **Developed** RESTful Backend Services using Python (Asyncio Sanic), **Redis** and **MongoDB** using **asynchronous** and **parallel programming** design patterns.
- **Designed** Real time communications API using web sockets and Kafka with high traffic.
- **Designed** scalable **Microservices** that includes low latency communication between API Gateway, Service Registry, Nginx, Zookeeper, Redis, Docker Swarm and Service Discovery Agent.
- **Developed NLP** predictions using **tf-idf** models and **knn** algorithm.
- Reduced the overall execution time of the **data pipelines** to **1/3rd,** with the same amount of resources.
- **Prioritized** product roadmaps with developers aligning product vision reaching delivery months before forecast.

Tata Consultancy Services, Kolkata, Nov 2010 - October 2018
Senior Software Engineer/Technical Lead

- **Designed and Developed** the **Warehouse Management Automation System** for a leading UK Retailer which increased the business efficiency by **30%** while leading the backend engineering team to deliver the cloud-based product across **600+** retail stores in UK.
- Worked with the client CTO and Architect on the product roadmap for Backend Engineering
- **Spearheaded** the clients global cross functional Backend Engineering team for their **Store App and Logistics Management System** that improved the company's ability to analyze and manage store logistics by **20%** and grew sales revenue by a huge amount, months ahead of plan.
- **Redesigned** the Consumer Portal for mobile enhancements for a leading water and electricity provider in Canada which resulted in **80%** revenue increase and manifold increase in the consumer

base.

- **Developed** a data analysis dashboard tool for the Utility provider in Canada while working closely with the global cross functional team.
- **Led** the Internal Utilities COE Solution group and created multiple solutions on Oracle Fusion Middleware which resulted in acquiring multiple high revenue projects by the horizontal
- **Front lined** the project design development and optimization of a leading Government Utility provider with Oracle Middleware and WebCenter ADF which resulted in achieving **60%** profit in their system.

Education

West Bengal University of Technology, Kolkata
Bachelor of Technology, Electronics and Communications **(8.6)** July 2006 - July 2010