REET NANDY

reet.nandy@nyu.edu | reetnandy.com | github.com/techpertz | linkedin.com/in/reetnandy | +1(518)9306116 | Willing to Relocate

PROFESSIONAL EXPERIENCE

Founding Software Engineer June 2025 – Present
Sentrion AI New York City, USA

As the Startup's **first technical hire**, I single-handedly architected, developed, and deployed the company's entire technical infrastructure and product suite from the ground up and increased the MRR by 2.5x under three months.

- Architected a high-performance data infrastructure from scratch for 60M+ rows, pairing PostgreSQL with RDS Proxy for connection pooling and Elasticsearch for sub-200ms search, and integrated an AWS Glue ETL pipeline to migrate 200M+ (1M+ daily) records with zero downtime.
- Engineered a multi-agent AI system using LangGraph to orchestrate AWS Bedrock models, automating complex user workflows and embedding intelligence directly into the core product (SaaS + Data Platform).
- Developed 2 fault-tolerant FastAPI microservices and a multiple RabbitMQ queues with priority routing to decouple SaaS, batch, and analytical workloads, eliminating $5k/month in vendor costs and ensuring 99.9% reliability with a centralised logging system with Better-Stack.
- Optimised cost for AWS, reducing monthly cloud spend by $3k through autoscaling, rightsizing, and event-driven (S3) daily data ingestion.
- Architected the platform's unified IAM and monetisation core, engineering a secure, tiered billing system by integrating Clerk (authentication), a custom JWT-based API credit ledger, and Stripe API.
- Served as the 1:1 technical POC for enterprise clients, translating complex business needs into engineering roadmaps to drive customer success.

Backend Development Intern (Machine Learning) June 2024 – December 2024
Mobility Intelligence New York City, USA

- Built a real-time price prediction model using regression and Kalman filtering, achieving less than 5% error on a 90-day forecast.
- Deployed ML inference FastAPI microservices with Celery, Redis, and autoscaling on AWS EKS, sustaining 150k+ requests with 99.9% uptime.
- Automated ETL pipelines in Airflow to process 15M+ daily records, reducing manual effort and enabling downstream analytics.
- Set up Prometheus + Grafana with SLIs and alerting, cutting MTTD by 60% and boosting incident response efficiency.

Software Engineering Intern (R&D) January 2023 – June 2023
Defence Research & Development Organisation India

- Engineered multithreaded architecture for real-time LiDAR processing, handling 50K data points/sec (97% accuracy).
- Implemented Redis-based geospatial caching over PostgreSQL/PostGIS, reducing GPS query latency from 1000ms to 150ms.
- Developed an ETL pipeline using memory-efficient streaming, processing 12GB/min while reducing memory usage by 60%.

Software Engineering Intern April 2022 – December 2022
Solar Industries India Ltd India

- Led a team of 5 to automate workflows, delivering 5 production-ready Django systems that standardised 80% of manual processes.
- Reduced API latency by 25% and integrated distributed tracing with Jaeger and Grafana, enabling real-time debugging.
- Designed a partitioned Kafka pipeline with scalable consumer groups, processing 2.5M+ rows/sec using Redis with 100K+ daily requests.

PROJECTS

[AI Agent] **GraphRAG - LLM Document Compliance** (Github) *April 2025*

- Launched an Agentic SaaS with NextJS for real-time document edit, approval, and audit reports via Graph-based RAG and LLM.
- Implemented a GraphRAG and PDF parser in Python from scratch for unstructured PDFs using PDFMiner & Tesseract (OCR).
- Executed semantic chunking + NLP NER + BART-CNN summarisation, achieving 70% relation extraction accuracy.
- Synthesised hybrid retrieval (Vector + Graph + metadata), boosting compliance accuracy to 90%.

[Python / Core] **Hierarchical Vector Database from Scratch** (Github) March 2025

- Built embedding database (library - document - chunk) with async collection mutexes; 12K ops/sec at <0.1% conflicts.
- Added 3 indexing algorithms (LinearScan/KD-Tree/LSH) for vector search on 10M vectors in 18ms.
- Led Kubernetes and Helm charts deployment along with a custom-made CLI toolkit, reducing onboarding complexity by 100%.

SKILLS

Frontend + Backend: Python (FastAPI, Django), TypeScript (Next.js), Tailwind CSS, Node.js (Express.js), Java (Spring Boot)
AI / LLM: Transformer (GPT, BERT, LLaMA), Graph + RAG, Fine-tuning, Embeddings, Quantization, Multi-Agent
Cloud + DevOps: AWS (EC2, Lambda, S3, SageMaker, Prometheus), GCP, Kubernetes, Docker, Jaeger, CI/CD, ELK, Microservices
Data + Pipeline: PostgreSQL, MongoDB, Vector DBs, Redis, Kafka, Airflow, ETL / ELT, REST, gRPC, Agile

EDUCATION

New York University – New York City, USA September 2023 - May 2025
Master of Science, Computer Science (MS) | *Merit Scholarship Recipient* GPA: 3.7/4.0

- *Relevant Coursework*: Data Structures and Algorithms, Cloud Computing, Machine Learning, Big Data Analytics, Database Systems, Software Engineering, Object-oriented Design, Probability and Statistics
- *Graduate Teaching Assistant*: CS GY 6233 - Operating System, CSCI UA 0310 - Algorithms

Manipal University Jaipur – India July 2019 - May 2023
Bachelor of Technology, Computer Science (B. Tech) GPA: 3.8/4.0