REET NANDY

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

SUMMARY

Software + AI Developer with 3 years of experience across 6 internships in Full Stack Development and **Distributed Systems**. Skilled in building resilient microservices, event-driven architectures, and zero-downtime deployments for cloud-native applications (AWS).

Currently Building Projects involving Multi-Modal AI Agents with LLMs, vision models, and Advanced RAG pipelines.

EDUCATION

New York University, Tandon School of Engineering – 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

New York University, Stern School of Business – New York City, USA — Jan 2025 - May 2025
Business Strategy *under Prof Adam Brandenburger* — GPA: 4.0/4.0

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

SKILLS

Languages: Python, JavaScript, TypeScript, SQL, C/C++, Java, HTML / CSS, Bash
Frontend + Backend: React, Next.js, Tailwind CSS, Django, FastAPI, Node.js, Express.js, Spring Boot, Microservices
AI / LLM: Transformer (GPT, BERT, LLaMA), Graph + RAG, Fine-tuning, Embeddings, Quantization, Multi-Agent
Cloud + DevOps: AWS (EC2, S3, Lambda), GCP, Kubernetes, Docker, Jaeger, Prometheus, Grafana, CI/CD, ELK Stack
Databases + APIs: PostgreSQL, MongoDB, Redis, Airflow, Kafka, RabbitMQ, REST, Socket.io, gRPC, OAuth2, Agile

PROFESSIONAL EXPERIENCE

Mobility Intelligence — New York City, USA
FullStack Development Intern — June 2024 – December 2024

- Built a real-time price prediction system using regression and Kalman filtering, achieving less than 5% error on 90-day forecasts.
- Designed a FastAPI backend with Celery and Redis, handling 150k requests daily with 99.9% uptime and sub-500ms P95 latency.
- Scheduled Airflow DAGs managing ETL pipelines processing 15M+ daily records from PostgreSQL into analytics-ready stores.
- Configured Prometheus + Grafana with SLIs and alerting rules, reduced MTTD by 60% and improved response workflows.
- Deployed microservices in AWS using Kubernetes with Helm charts, rolling updates, and horizontal pod autoscaling, reducing downtime during deployments by 80% and enabling seamless CI/CD.

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

- 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 ETL pipeline using memory-efficient streaming, processing 12GB/min while reducing memory usage by 60%.

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

- Led a team of 5 to automate workflows, delivering 5 Django microservices that standardized 80% of manual processes.
- Reduced API latency by 25% and integrated distributed tracing with Jaeger, enabling real-time debugging.
- Designed a partitioned Kafka pipeline with scalable consumer groups, processing 2.5M+ rows/sec using Redis caching and Cassandra-backed storage.

PROJECTS

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

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

Next Page —>

[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 Helm deployment along with custom made CLI toolkit, reducing onboarding complexity by 100%.

[AWS / Fullstack] **AI-Fitness Analytics Dashboard** (Github) April 2024

- Deployed Django API on AWS Elastic Beanstalk with Google Fit integration, configured auto-scaling groups + health probes.
- Orchestrated pipelines with Lambda-SageMaker, deploying KNN models for recommendations at 92% accuracy.
- Coordinated event-driven metrics processing via SNS/SQS, achieving 800ms p95 latency for real-time health data.

[AI Workflow] **Multi-Platform LLM Task Orchestrator** (Github) December 2024

- Created Flask API with Gmail/Slack/Trello via OAuth2, automating task extraction and priority classification with OpenAI.
- Configured MongoDB RAG system with vector indexing for context memory management via chat history tracking.
- Automated cross-platform workflow triggers by combining LLM with rule-based logic, reducing manual task handling by 60%.

[Java / Fullstack] **Real-Time Collaborative Whiteboard** (Github) December 2024

- Built low-latency collaboration using Spring Boot and Swing, achieving <150ms sync for concurrent users via binary compression.
- Implemented vector operations using operational transformation, resolving 98% conflicts in real-time updates.
- Executed socket programming with PostgreSQL and JSONB storage, achieving 85% network overhead reduction.