Keshav Jha

keshavsde@gmail.com | +91 9599015933 | linkedin.com/in/therealsaitama | github.com/therealsaitama

Education

Delhi Technical Campus (GGSIPU), B.Tech in Computer Science — Nov 2022 – Expected Nov 2026

- **Coursework:** Data Structures & Algorithms, Operating Systems, Computer Networks (TCP/IP), Distributed Systems, Database Management Systems, Information Retrieval, Machine Learning

Experience

Lead AI Engineer & DevOps Architect, SNK Consultancy – New Delhi, India — Aug 2024 – Present

- Spearheading the **AI/ML strategy** and **cloud infrastructure**; architecting scalable pipelines to serve machine learning models for high-throughput consultancy operations.
- Managing the complete DevOps lifecycle; implemented **Docker** containerization and **GitHub Actions** CI/CD pipelines to automate testing and reduce deployment latency by 40%.
- Optimizing model inference and backend performance; established real-time monitoring and alerting stacks to ensure zero downtime for critical business services.

Co-founder & Lead Backend Engineer, She&Soul – New Delhi, India — Jan 2024 – Aug 2024

- Architected and implemented the platform backend using **FastAPI**, **PostgreSQL**, and **Redis**; designed REST APIs, JWT auth, RBAC, and a multi-tenant data model.
- Built workflows for payments, orders, and notifications (FCM/SMS); added rate limiting, request tracing, and a zero-downtime migration strategy for PostgreSQL schema changes.

Projects

PyDB-Core — Relational Database Kernel | *Python, Databases, Systems* — June 2025 – Aug 2025

- Engineered a small persistent relational store in pure Python; implemented **B+ tree** indexing and an **LRU** buffer pool for on-disk page management.
- Built a basic **SQL parser** (AST construction with simple predicate pushdown) to minimize full-table scans; experimented with snapshot-style reads to keep queries responsive during writes.
- Added a simple write-ahead log (WAL) and crash-recovery routine to reduce the risk of data loss during write-heavy workloads.

Raft-Cluster — Distributed Consensus System | *Python, asyncio, TCP/IP* — Aug 2025 – Oct 2025

- Developed a fault-tolerant key-value store implementing the **Raft** consensus algorithm (leader election and log replication) to keep replicas in sync across a 3-node cluster.
- Constructed a custom **RPC layer over TCP** using `asyncio` with periodic heartbeats; in failure tests, leader re-election completed within hundreds of milliseconds while keeping writes consistent across nodes.

Neuro-Stream — Distributed Parameter Server | *Python, gRPC, ML Infra* — Oct 2025 – Present

- Built a parameter-server prototype in Python/**gRPC**, coordinating N workers sending asynchronous gradient updates during training.
- Experimented with bounded-delay asynchronous updates to study stale-gradient effects on convergence; used simple **Int8** quantization on ResNet50 parameter payloads to cut gradient payload size by about 75% in tests.

Achievements

Competitive Programming: LeetCode Knight (Top 3%, Max 1938) • CodeChef 5 Star (Max 2077, India Rank 540) • GeeksForGeeks Global Rank 41 (Max 2180) • Codeforces Top 0.6% (Global Rank 327) • 2000+ DSA Problems Solved

Technologies

Languages: Python, C++, Go, SQL, JavaScript
Systems & Networking: Linux/UNIX, TCP/IP, Sockets (epoll), WebSockets, gRPC, HTTP/REST
Frameworks & Tools: FastAPI, Flask, PostgreSQL, Redis, Docker, Nginx, Git, GitHub Actions, Prometheus, Grafana
ML Libraries: PyTorch, ONNX Runtime, NumPy, scikit-learn