

SRI HARSHA PATALLAPALLI

+918110850621 ◇ Hyderabad ◇ sriharsha.py@gmail.com ◇ [linkedin.com/in/sriharsha-py](https://www.linkedin.com/in/sriharsha-py) ◇ github.com/sriharshapy

EDUCATION

Northeastern University, Boston, MA
Master's in Artificial Intelligence. GPA: 3.833/4

Sept 2023 - Dec 2025

Amrita School of Engineering, Coimbatore, India
B-Tech in Computer Science and Engineering. CGPA: 8.24/10

Jun 2015 - May 2019

EXPERIENCE

Career Transition & Relocation

Hyderabad, India | Jan 2026 - Present

- Relocated from the United States to Hyderabad after an MS at Northeastern (Dec 2025), continuing open source work on Crucible.

[Dextr.ai](#) | Founding Software Engineer

Bay Area, CA | Dec 2024 - Jun 2025

- Designed the REST backend and middleware on AWS EKS for 10K concurrent guests, with sub 100ms latency on in house paths.
- Architected multi channel workflows in Python with asyncio and FastAPI around a shared conversation state store, keeping guest sessions persistent across voice and SMS, and cutting response time from 4.2 hours to under 10 seconds.
- Owned code reviews and the GitHub Actions CI/CD pipeline for Terraform managed EKS deploys, covering source control management, automated build processes, and testing that kept a 24/7 multi tenant platform and its auth system releasable on every merge.
- Instrumented Prometheus and Grafana observability over counters, errors, and EKS health, and hardened the third party voice integration with timeouts, bounded retries, and a circuit breaking dead letter queue for zero missed calls across 24/7 coverage.
- Made the agents fail closed, declining out of scope requests instead of guessing and writing structured failure records that an hourly triage agent escalated to hotel staff, cutting in stay response from 47 minutes to under a minute.

Lowe's | Software Engineer (Data Platform)

Bangalore, India | Jun 2021 - Jun 2023

- Shipped Java Spring Boot microservices on MySQL across the full software development life cycle, owning API design, JVM tuning, and reliability for over 10 teams.
- Delivered a Kafka and ClickHouse streaming service that ingests 10K+ clickstream events/sec from web platforms and backs real time customer behavior analytics and personalization at sub second query latency.
- Developed Teradata SQL and Java workflows on Hadoop via Airflow that track roughly \$300M in annual Pro discounts, which surfaced margin leakage and shaped pricing and promotion decisions for retail finance.
- Owned and operated the weekend on call rotation, maintaining reliability by hardening the Bash jobs behind escalations.
- Drove a fault tolerant Spark system in Scala and PySpark on Hive and Airflow that processes 2TB of data within a 6 hour SLA.

GE Healthcare | Software Engineer

Bangalore, India | Aug 2019 - Jun 2021

- Delivered medical equipment tracking on AWS EKS with Kafka, Spring Boot, and Elasticsearch at 70M+ records/day under 2 s lag.
- Hardened ingestion with retries, dedupe, ordering, schema evolution, and backpressure, and loss fell below 0.01%.
- Centralized logging and observability pipelines with Elasticsearch and alerts, which improved SLA and incident response 75%.
- Migrated a petabyte scale on premise distributed system to AWS on Redshift, S3, Glue, and PySpark at 80% lower cost.
- Built a config driven SQL emitter in C++ that builds an AST to derive table and column relationships and generate schema DDL.

PLATFORM AND DEVELOPER TOOLING PROJECTS

Crucible | Open-Source Correctness and Timing Harness for GPU Kernels ([repo](#)) ([docs](#))

- Authored Crucible, an open source harness verifying GPU kernels against an fp64 CPU reference at 95% confidence.
- Caught real defects with it: tail dropping and contiguity assumptions failed 5 of 16 and 15 of 16 generated shapes, with each failure minimizing to a reproducible seed and the tool shipping as a zero config pytest plugin.

GrepIt | GPU Accelerated Pattern Search Library ([repo](#))

- Built a pattern search library in C++17 and CUDA that scans 10GB+ files by chunking them through GPU memory automatically, exposing a single API call for both in memory buffers and files, with regex wildcards, quantifiers, and character classes.
- Optimized CUDA kernels for Ampere with 256 thread blocks and shared memory caching to cut redundant global memory reads.

TECHNICAL SKILLS

Languages & Frameworks: Java, Spring Boot, Python, C++, SQL, Bash, Git, Jenkins, Docker, Kubernetes, Linux

Distributed & Streaming: Kafka, Spark (PySpark), ClickHouse, Hive, Hadoop, Airflow, event-driven architecture, idempotency

Services & Datastores: REST, gRPC, microservices, MySQL, PostgreSQL, Oracle, DynamoDB, Redis, Elasticsearch

Cloud & Operations: AWS (EKS, S3, Glue), Terraform, Helm, CI/CD, Prometheus, Grafana, observability, reliability, on-call

CERTIFICATIONS

Confluent Certified Developer: Apache Kafka | **AWS Certified Data Analytics:** Specialty | **NVIDIA-Certified Associate:** GenAI LLMs

AWARDS AND RECOGNITION

- Best in Enterprise Computer Vision**, HP & NVIDIA Developer Challenge (\$5,000 prize, 1000+ participants).
- Winner**, GE Digital Industrial IIoT Hackathon (INR 100,000 prize, 180+ teams competed).
- 2nd Runner Up**, K-GIS Hackathon at the Bengaluru Tech Summit Global Hack (INR 100,000 prize).
- Recognized as the best mentor** at Lowe's for training and onboarding new recruits.