Raghav Kachroo

(858)-241-1760 | rkachroo@ucsd.edu | linkedin.com/raghavkachroo | github.com

Education

University of California, San Diego Sep 2024 - Mar 2026
Master of Science in Data Science (Artificial Intelligence & Machine Learning)

Indraprastha Institute of Information Technology, Delhi Sep 2022 - Sep 2023
Post Graduate Diploma in Data Science & Artificial Intelligence

University of Delhi Jul 2018 - Jun 2021
Bachelor of Management Studies

Experience

Amazon Jun 2025 - Sep 2025
Software Development Engineer Intern Bellevue, WA

- Built a distributed log indexing and query service over **42M+** production log entries, implementing parallelized indexed queries to reduce triage latency from **15+ minutes to under 45 seconds** during incidents.
- Designed a workflow automation layer using AWS Step Functions and Lambda to automate on-call SOPs, saving **12+ engineer-hours per week**.
- Published the log access layer to diagnostic tooling via MCP, optimizing caching and query batching to sustain **sub-2s** response times during outage triage.

Aark Global Apr 2023 - Sep 2024
Software Developer, AI/ML Delhi, India

- Built a document ingestion pipeline processing **18,000+ pages/day** by distributing tasks through Azure Queue Storage to VM workers, preventing processing backlogs during traffic spikes.
- Implemented a read/write routing layer directing reads to Cosmos DB and writes to MongoDB, delivering **sub-100ms P95** response times.
- Built an OCR-backed search pipeline that converted unstructured scanned PDFs into indexed Elasticsearch documents, enabling **sub-180ms** query latency.
- Reduced infrastructure spend by **32%** by profiling VM utilization and database throughput, adjusting worker counts and database configurations through controlled A/B tests.

Concentrix Jun 2022 - Mar 2023
Data Engineer Gurugram, India

- Replaced sequential web scrapers with distributed ingestion jobs orchestrated in Airflow and streamed through Kafka, increasing data ingestion throughput by **60%**.
- Migrated downstream aggregations from batch jobs to Kafka streaming stages, reducing analytics data freshness lag from **3 days to 6 hours**.
- Added Spark data validation checks and structured error logging to ingestion workflows, reducing mean time to resolution (MTTR) from **6 hours to 2 hours**.

Projects & Research

TRAI: AI Mobile App to Reduce the Gap Between Triage and Care | *IEEE* July 2025

- Built and deployed a clinical decision support service exposing ML inference via REST APIs, reducing model inference time from **16s to 5s** through caching and request batching optimizations.

Generalized Prediction of Shock in Intensive Care Units using Deep Learning | *medRxiv* Jul 2021

- Developed and evaluated deep learning models for ICU time-series shock prediction, evaluating generalization across age groups and clinical settings.

VoiceCode: Multi-Agent Inference & PR Automation Platform | *GitHub* Jan 2026

- Architected a multi-agent LLM inference system converting natural language into pull requests, orchestrating sandboxed execution with retrieval-augmented generation, achieving **sub-60s** average time-to-PR.

Hao AI Lab – Diffusion Model Optimization Jan 2026 - Present

- Profiling diffusion model training and sampling pipelines to improve GPU utilization and reduce inference latency.

Technical Skills

Programming: Python, Java, SQL, C++
AI & LLM Systems: PyTorch, Hugging Face Transformers, LangChain, LangGraph, RAG Pipelines, OpenAI APIs
Distributed & Data Systems: Kafka, Spark, Airflow, MongoDB,Cosmos DB, Elasticsearch, Azure Queue Storage
Cloud & Infrastructure: AWS (EC2, S3, Lambda), Azure (VMs, Queue Storage), GCP, Docker, GitHub Actions, Linux
Developer Tooling: REST APIs, FastAPI, Redis, MLflow, Git