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

- Designed and shipped a log query service across **42M+ distributed records**, reducing query latency from **15+ minutes to <45 seconds** and enabling SREs to triage alerts, validate rollbacks, and debug failures during outages.
- Integrated the service into automated incident workflows using AWS Step Functions, replacing manual SOP execution and saving on-call engineers **12+ hours per week**.
- Productized the log query service as a reusable internal SDK adopted across 7 teams, standardizing how services query logs and trigger incident-response actions.
- Exposed the log access layer to troubleshooting tools via MCP, optimizing for **<2s response times** during interactive debugging sessions.

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

- Designed a distributed document processing pipeline handling **18,000+ pages/day** using Azure Queue Storage to distribute work across VM fleets, preventing ingestion backlogs during traffic spikes.
- Implemented a hybrid data serving layer delivering **sub-100ms P95 response times** by routing latency-critical reads through Cosmos DB while retaining MongoDB for flexible writes.
- Built a search system over scanned documents by piping Tesseract OCR output into Elasticsearch, achieving **<180ms query latency**.
- Improved pipeline reliability by making queue workers idempotent and isolating stage failures, cutting average recovery time from **4 hours to 90 minutes** during backlog spikes.

Concentrix Jun 2022 - Mar 2023
Data Engineer Gurugram, India

- Designed and operated scalable data ingestion pipelines for high-volume social and e-commerce data, replacing legacy scrapers with parallelized Airflow and Kafka workflows to increase monitored brand coverage by **60%**.
- Converted downstream analytics from batch jobs into streaming pipeline stages, reducing end-to-end data availability from **3 days to 6 hours** in production.
- Added Spark-based data quality checks and structured error logging to ingestion pipelines, cutting time to diagnose and fix data failures from **6 hours to 2 hours**.

Research & Projects

TRAI: AI Mobile App to Reduce the Gap Between Triage and Care | *https://ieeexplore.ieee.org/document/11252667*

- Built and deployed a clinical decision support service exposing ML inference via REST APIs, integrating structured EHR data into a latency-sensitive request/response workflow.
- Profiled end-to-end inference latency across orchestration and model execution paths, reducing response time from **16s to 5s** through caching, request routing, and batching tradeoffs.

Autonomous GPU Rental Agent | *GitHub*

- Designed a long-running infrastructure orchestration system to provision GPUs, execute jobs, and handle retries across unreliable external vendors.
- Implemented a MongoDB-backed state machine to persist workflow state across stateless service restarts, enabling crash-safe recovery for multi-hour execution paths.

Technical Skills

Languages: Python, Java, SQL
Cloud & Infrastructure: Distributed Systems, Cloud Services (AWS, Azure, GCP), Docker, CI/CD, Monitoring
Data Systems: MongoDB, Cosmos DB, Elasticsearch, Apache Kafka, ETL Pipelines, Data Modeling
ML Systems: PyTorch, Inference Pipelines, Model Evaluation, Retrieval-Augmented Systems
Developer Tooling: REST APIs, Workflow Orchestration, Model Context Protocol (MCP), Agent Frameworks