Rahul Ganesan

303-847-9894 | g.rahulganesan107@gmail.com | linkedin.com/in/rahul-ganesan-cu/ | Boulder, CO

Professional Experience

Frazier Simplex Machine Company — Washington, PA
Data Engineer Intern — *Jun. 2025 – Aug. 2025*

- Independently architected and built an OCR based document ingestion pipeline with Spark and AWS RDS. Designed ETL workflows and optimized PostgreSQL schema to process 100GB of CAD engineering drawings.
- Automated digitization of 5000 drawings, reducing processing time by 80% and eliminating manual effort. Designed real time Tableau dashboards from transformed data that accelerated planning and vendor selection workflows.

Projects

MOSAIC — Multi-Modal Orchestrated System for AI Chat — Jan. 2026 - Feb. 2026
Agentic system for voice/text information retrieval across web and knowledge bases

- Architected a multi-modal agentic system using FastAPI with LangGraph, implementing multi step tool calling via Model Context Protocol across Notion APIs and web search tools. Designed explicit state tracking, tool routing logic, and fallback recovery to ensure robust query decomposition and retrieval consistency.
- Improved end-to-end agent reliability and tool success rate by 22% while maintaining 3–5s response latency by refining tool schemas, prompt hierarchies, and routing heuristics. Reduced cognitive switching cost for researchers by unifying search, documentation, and retrieval workflows into a single conversational interface.

LIFT - LLM Inference with Fast Throughput — Dec. 2025 - Feb. 2026
GPU-backed Aysnc LLM Serving with Micro Batching, Redis Queues and observability

- Architected a GPU based LLM serving system using FastAPI and vLLM with async request handling, Redis queues for backpressure control, and micro batching to maximize GPU utilization under concurrent multi user traffic.
- Improved latency and inference efficiency under Locust-driven load tests by implementing Redis caching and request deduplication, reducing query latency from 8.3 s to 5 ms. Integrated Prometheus + Grafana observability to monitor latency, cache utilization and overload conditions in a production-style LLM serving environment.

BENCH — Benchmarking Engine for Neural Code Helpers — Jan. 2026 – Mar. 2026
End-to-end pipeline for measuring LLM codegen cost, latency, and correctness

- Engineered an EvalPlus-based evaluation harness around StarCoder2 models on the 164 problem HumanEval(+ via EvalPlus) split, wiring model calls into an agent-coding-bench-style pipeline that produced pass@1 execution traces under strict unit-test checking.
- Instrumented and benchmarked end-to-end cost and latency to compare model configurations, logging 1.5s mean codegen latency and 40k tokens across the run, and hardened evaluation for evaluation reliability by normalizing outputs to reduce formatting induced test failures, enabling more consistent benchmarking across model configs.

Event Driven Music Separation Platform on Kubernetes — Sep. 2025 - Dec. 2025
Cloud SaaS with Redis Caching, and Autoscaling for Asynchronous Process

- Architected an event driven microservices platform on Kubernetes and Google Cloud using Docker, Redis caching/rate limiting, and Demucs for audio source separation, orchestrating asynchronous job queues through stateless worker pods and REST based ingestion services for scalable concurrent audio processing.
- Improved processing throughput by 20% through parallelized worker execution and Kubernetes Horizontal Pod Autoscaling (HPA), enabling reliable handling of concurrent uploads under bursty streaming-style workloads while reducing end-to-end preprocessing latency.

Education

University of Colorado Boulder — Boulder, CO
Master of Science, Data Science — *Aug. 2024 – May 2026*

Skills

Big Data Tech: Hadoop, Spark, Airflow, Kafka, Docker, Kubernetes
Techniques & Software: n8n, FastAPI, React, Jenkins, Node.js, Prometheus