Rahul Surya

Edinburgh, UK · +44 7742 977485 · rahulsurya021@outlook.com · LinkedIn · GitHub

SUMMARY

MLOps and Platform Engineer with production ML deployment experience at ISRO and Microsoft Azure certifications (DevOps Engineer Expert, Developer Associate), currently completing an MSc in High-Performance Computing with Data Science at the University of Edinburgh. Proficient in Docker, Kubernetes, CI/CD pipelines, FastAPI, and cloud infrastructure. Proven ability to operationalise ML models at scale and build reliable, automated software delivery platforms. Seeking MLOps and Platform Engineer roles in the UK.

SKILLS

- **MLOps & Serving:** MLflow, FastAPI, Docker, Kubernetes, CI/CD (Azure Pipelines, GitHub Actions), Model Registry, A/B Testing.
- **Cloud:** Microsoft Azure (DevOps Expert, Developer Associate – Certified), AWS basics, GCP basics.
- **Infrastructure as Code:** Terraform, Helm, Azure Resource Manager.
- **Languages:** Python, Go, C++, Java, SQL, Bash.
- **ML Frameworks:** PyTorch, TensorFlow, Hugging Face, Scikit-learn, CUDA.
- **Data:** Apache Kafka, Apache Spark, PostgreSQL, MongoDB, Redis, Elasticsearch.

PROFESSIONAL EXPERIENCE

Machine Learning Engineer Intern · ISRO – National Remote Sensing Centre — Oct 2023 – Jan 2024

- Deployed a production **ConvLSTM-Seq2Seq** lightning prediction system in PyTorch with 92% accuracy, integrating real-time inference via REST APIs into a live-serving React.js dashboard.
- Optimised deep learning data pipelines on **500GB+ of WRF simulation data**, cutting model training latency by **35%** using VAE compression and efficient parallel data loading.
- Served live predictions to **15+ meteorologists** daily with zero-downtime model updates.

Software Engineer Intern · Develoscope Software Solutions — Jun 2023 – Sep 2023

- Delivered 4 production Java web applications (JSP, Apache Tomcat) with end-to-end CI/CD and REST API integrations, increasing client satisfaction by **25%**.

Data Science Intern · Clustor Computing — Jun 2023 – Sep 2023

- Built and deployed ML models for financial forecasting achieving 94% directional accuracy.

PROJECTS

LLM Inference Engine Optimisation · *Python, FastAPI, CUDA, PyTorch, Docker* — *MLOps Project*

- Engineered a scalable, containerised inference server for transformer models with GPU memory management, continuous batching, and quantisation-aware serving via FastAPI.
- Designed for horizontal scaling with Kubernetes, supporting automated load-based scaling.

Real-Time Distributed Log Analytics System · *Kafka, Spark, Docker, Kubernetes, Elasticsearch*

- Built a containerised real-time log processing pipeline (Kafka + Spark Streaming) deployed with Docker and Kubernetes, with Kibana dashboards enabling **70% faster** issue resolution.

Distributed ETL & Caching Pipeline · *Python, PostgreSQL, MongoDB, Redis*

- Architected a high-throughput pipeline ingesting 1GB+ of financial data into a hybrid SQL/NoSQL schema with Redis caching, reducing query latency to **<10ms**.

EDUCATION

University of Edinburgh – *MSc High-Performance Computing with Data Science* — Sep 2025 – Aug 2026
Coursework: Distributed Systems, High Performance Data Analytics, GPU Programming (CUDA), Machine Learning, Reinforcement Learning.

G H Raisoni Institute of Engineering – *B.Tech Artificial Intelligence (First Class with Distinction)* — Aug 2020 – Jul 2024

CERTIFICATIONS

- **Microsoft: Azure DevOps Engineer Expert**, **Azure Developer Associate**, Azure Fabric Analytics Engineer, AI Fundamentals, Data Fundamentals.