OMKAR UBALE

AI Systems Engineer | Scalable RL & Distributed ML Infra | AWS, Java, Spark, Kubernetes
+1 332 699 1882 | oubale@gmail.com | linkedin.com/in/omkarubale | omkarubale.com | GitHub: github.com/omkarubale

EDUCATION

Northeastern University, Boston, MA - M. S. in Computer Science (GPA: 3.95/ 4) Sep '22 – Dec '24

Relevant Coursework: Algorithms, Machine Learning, Parallel Distributed Systems (Hadoop, Spark), Artificial Intelligence

Vellore Institute of Technology, Vellore, India - B.Tech. Electronics & Communications Aug '14 – May '18

WORK EXPERIENCE

Amazon - AGI SF Lab (Nova Act) *Boston, MA*

Software Development Engineer 2 Jan '25 – Present

- Architected **multi-region RL environment hosting platform** for Nova Act's foundational model training (100k+ concurrent environments across 6 AWS regions, 47k TPS, 99.99% uptime, 20 ms latency) using Java, AWS CDK, and DynamoDB — enabling large-scale policy training for AGI experiments
- Automated EC2 AMI image creation for **10 vendor-provided environments** using **AWS Image Builder**, **AWS CDK** and **Docker Compose**, reducing provisioning time **30 min → 2 min** per image and standardizing environment specification
- Designed and built backend for internal **Nova Act playground** serving **3k+ concurrent sessions** with **sub-second latency** using Nova Act SDK, Bedrock AgentCore Runtime, and Remote Browser over Web Sockets (Chrome DevTools Protocol)
- Eliminated recurring error code **424 errors (5% of all requests → 0%)** in Nova Act Research Preview by re-architecting core backend services, improving reliability and user experience, and improving **availability from 85% to 99.5%**

Sway AI *Burlington, MA*

Back-end Software Engineer Co-op Jun '23 – Dec '23

- Delivered model interpretability tooling (**Partial Dependence Plots, What-If Analyzer**) for 20+ production ML models, **cutting evaluation cycles from 2h → 30m** and improving transparency for enterprise clients
- Deployed Kubernetes observability stack (**Prometheus + Grafana** on **AWS EKS** cluster) to track model inference performance (CPU, memory, error rates, replica counts)
- Optimized Django APIs and AWS Lambda configs, reducing **p50 latency from 17s → 0.8s** via query optimization + provisioned concurrency

Rare Crew *Bengaluru, India & Belgrade, Serbia*

Senior Software Engineer Aug '18 – Jun '22

- Led engineering for ERP product built using .NET and Angular with **50k+ active users** and **250M+ unique access configs**, building core modules and integrations (SAP, Power BI) that improved asset visibility for 14k+ assets, reducing reporting time from **60 hours to 4 hours** per week
- Built Azure CI/CD pipelines reducing deployment time from 2 hours → 7 minutes; launched a **mentorship program for 50+ developers**, and co-led a 60-member team

PROJECTS

Retrieval Augmented Generation (RAG) Chatbot [GitHub] [Live Demo] Oct '25 - Present

- Built a RAG chatbot for e-commerce store using **LangChain, FAISS,** and **Groq**; implemented document ingestion, context aware chunking, and local embedding retrieval to ground LLM responses in store policies, reducing hallucination
- Built as a **fully local LLM inference pipeline** (Gemma-3-1B via Ollama) and remotely hosted using Groq using Streamlit UI, enabling offline and explainable Q&A — showcasing production-grade retrieval and prompt-engineering design

Clustering 1M Songs with Spark (K-means & Hierarchical) [GitHub] Dec '24

- Implemented two distributed K-means designs in Spark Scala; ran locally, pseudo-distributed, and on **AWS EMR**
- Observed **~2.04× speedup on 500k rows (3→5 workers)** via parallelized multi-K K-means; doubling data (**500k→1M, 5 workers**) scaled **~1.86×**.
- Observed **~34% runtime reduction** in hierarchical clustering using repartitioning + per-partition minima; documented shuffle trade-offs and **$O(n^2)$ memory pressure**.

TECHNICAL SKILLS

Languages: Java, Python, C#, Scala, SQL, TypeScript, C++, C, JavaScript, Bash
Frameworks/Libraries: Apache Spark, Hadoop, Kafka, Django, .NET Core, React.js, Angular, AWS CDK, JUnit, Node.js
Cloud & Infrastructure: Distributed Systems on Cloud, AWS, Kubernetes, Azure, Jenkins, Linux, Git, CI/CD
Databases: MySQL, PostgreSQL, MongoDB, SQL Server, Firebase Realtime Database

CERTIFICATIONS

AWS Certified Solution Architect - Associate, Amazon Web Services Aug '24