Ankit Suresh Raut

Fairfax, VA | 571-224-4890 | ankitraut9421@gmail.com | linkedin | github

PROFESSIONAL SUMMARY

Software Engineer and Cloud Infrastructure Lead with 3+ years of experience designing scalable AI, backend, and cloud-native systems. Proven record of delivering a 42% boost in reliability, sustaining 98% uptime, and achieving 70% faster deployments through CI/CD and serverless automation. Skilled in Java, Golang, Python, React, and AWS/Kubernetes, with expertise in distributed systems, DevOps practices, and performance optimization. Passionate about building fault-tolerant systems and driving measurable business impact.

PROFESSIONAL EXPERIENCE

YesTech Corp. **Feb 2025 - Present**
Full-Stack Engineer and Cloud Infrastructure Lead *Remote*

- Led backend services in Golang and Python, supporting 500+ AI image-generation workflows/month with scalable API design, observability, and typed SDKs.
- Architected and optimized AWS infrastructure (Athena, DynamoDB, S3, Lambda, Bedrock, CloudWatch, RDS), improving reliability by 42% and reducing spend by $900/month.
- Integrated LLMs, RAG, PyTorch, and OpenCV into AI pipelines, boosting image quality by 15% and increasing throughput by 20%.
- Built developer tooling in SwiftUI for log visualization, diagnostics, and test prompts; cut triage time per bug by 25-30%.
- Implemented blue/green deployments and automated rollbacks, reducing change-failure rate by 45%.
- Mentored a distributed team of 7 engineers, improving feature delivery with CI/CD and observability practices.

LTIMindtree Limited **Jun 2022 - Jan 2024**
Software Engineer P1 *Pune, India*

- Built and optimized REST/gRPC services, handling 2M+ monthly requests at 98% uptime with rate limiting, load balancing, and back-pressure.
- Integrated event streaming with Kafka/SNS/SQS, boosting system throughput by 35% and lowering latency via exactly-once/idempotent processing.
- Designed modular multi-tenant components for 30k+ daily transactions; introduced API versioning and contract testing.
- Applied Oracle Digital Assistant + AWS Lambda for conversational AI; reduced median response time by 2s, supporting 1,000 concurrent sessions.
- Implemented CI/CD with Docker, Jenkins, Terraform, cutting deployment errors by 40% and enabling canary rollouts.
- Authored design docs/RFCs, participated in cross-team reviews, and mentored junior engineers on system design and clean coding.

LTIMindtree Limited **Feb 2022 - May 2022**
Java Full-Stack Developer Intern *Pune, India*

- Developed full-stack features with Spring Boot, Angular, and PostgreSQL, contributing to 5 successful production releases.
- Designed optimized PostgreSQL schemas with Flyway migrations, improving p95 query latency by 120ms.
- Built hospital reception workflows processing 450+ daily appointments, reducing check-in time by 20%.
- Implemented CRUD services with in-memory caching, enabling 2,000+ concurrent users and lowering data-entry errors by 15%.

PROJECTS

SQL Query Chatbot

- Built a chatbot that translates natural-language queries to SQL with 92% accuracy, serving answers in <2s for 180 concurrent users.

Payments & Growth Platform

- Built Java microservices (REST/gRPC) processing 100k tx/day at 99.9% availability; reduced processing latency by 30 ms using idempotent workflows, retry/backoff, and event-driven messaging, with SLA/SLO/SLI and CloudWatch observability.

Air Traffic Control Voice Dashboard

- Transcribed controller audio to text with 95% accuracy; surfaced timelines/insights that cut manual review by 15 min/shift. Added scripts and metrics to evaluate models and pipeline health.

3D Gaussian Splat Segmentation (Meta's EgoLifter)

- Extended a PyTorch pipeline with SAM/SAM2 and contrastive learning; explored native 3DGS attributes with MLPs; achieved 25 dB PSNR in cloud GPU experiments.

Traffic Sign Detection and Recognition

- Real-time inference at 30 FPS with 92% accuracy; frame it as latency-sensitive serving, CPU/memory efficiency, and profiling under load rather than CV theory

Fake Content Detection in Text and Images

- Built ML inference pipelines reaching 92% classification accuracy and 50ms per-image latency; packaged models as reusable modules with tests and a simple evaluation harness.

SKILLS

- **Languages**: Java, Python, Golang, C#, JavaScript, SQL
- **Frameworks**: Spring Boot, FastAPI, React, Angular, Material-UI, SwiftUI
- **Databases**: PostgreSQL, MySQL, DynamoDB, RDS, MongoDB, SQL Server, Oracle DB
- **Developer Tools**: Git, Docker, Kubernetes, Terraform, Jenkins, GitHub Actions, VS Code, Py-Charm
- **Cloud Platforms**: AWS (EC2, S3, Lambda, API Gateway, RDS, DynamoDB, CloudWatch, Athena, Bedrock), GCP, Azure
- **DevOps / Practices**: CI/CD, Serverless, Agile Scrum, Object-Oriented Design, System Architecture, Observability, Troubleshooting
- **Testing**: JUnit, Cypress, PyTest, Postman, Testcontainers, XCTest
- **Machine Learning / AI**: PyTorch, scikit-learn, OpenCV, LLMs, RAG, Pandas, NumPy, Matplotlib

EDUCATION

George Mason University **Jan 2024 - Dec 2025**

Master of Science, Computer Science (AI/ML Concentration)

- **GPA:** 3.8/4.0

Pune University **Jun 2018 - Apr 2022**

Bachelor of Engineering, Computer Engineering

- **GPA:** 9.30 / 10

CERTIFICATIONS

- **Oracle Cloud Infrastructure 2023 AI Certified Foundations Associate:**Dec 2023
- **Oracle Cloud Infrastructure 2023 AI Certified Data Science Professional:**Dec 2023
- **Oracle Cloud Infrastructure 2023 Digital Assistant Professional:**Aug 2023
- **Java Programming Masterclass covering Java 11 and Java 17:**Udemy (Aug 2021)
- **AWS Certified Developer – Associate:**in progress

PUBLICATION

- Automated Traffic Sign Detection Using CNN.International Research Journal of Modernization in Engineering Technology and Science, Vol. 4, Issue 4, April 2022.