Shubham Asgaonkar

DevSecOps / MLOps Engineer

Mumbai
Model Town, Andheri West, Mumbai 400053
asgaonkarshubham70@gmail.com
9619739602
LinkedIn: ttps://www.linkedin.com/in/ shubham-asgaonkar/

DevSecOps & MLOps Engineer specializing in LLM deployment, scaling, and optimization. Experienced in building end-to-end MLOps/LLMOps pipelines, deploying GPU-accelerated workloads on Kubernetes, and integrating monitoring & observability solutions. Skilled in CI/CD for ML workflows, GPU/TPU optimization, quantization, batching, and secure model-serving. Adept at collaborating with ML research teams to operationalize fine-tuned and multi-agent LLM workflows.

Professional Experience

Associate Cloud Engineer *ONETURE TECHNOLOGIES , Mumbai | September 2024 - Present*

- Secured and optimized cloud environments across 4+ AWS accounts, improving security compliance scores from ~20% to over 85% using AWS Config, Inspector, and GuardDuty.
- Migrated infrastructure from public to private subnets, strengthening network isolation and compliance across environments.
- Deployed client-specific staging environments with Amazon Linux EC2 and RDS; configured Nginx reverse proxy for React (frontend) and Node.js (backend) apps with MongoDB/PostgreSQL on RDS/EC2.
- Containerized frontend and backend services using Docker Compose, integrated Fluentd for log scraping, and set up Prometheus + Grafana dashboards for observability and real-time monitoring.
- Built a production-grade MLOps pipeline for GPU-based image recognition models:

1. Containerized 15+ models with FastAPI, deployed on EKS GPU nodes (g4dn.xlarge, g4dn.2xlarge) using GPU time-slicing for efficient utilization.
2. Designed an automated retraining-to-deployment workflow: trained models stored in S3, versioned in CodeCommit, and deployed via Lambda-triggered CI/CD pipelines.
3. Implemented Lambda-based model versioning (extract → rename → maintain history) and automated container builds (Docker → ECR → EKS) using CodePipeline & GitLab for multi-environment deployments (DEV, UAT, PROD).
4. Optimized inference serving to handle 400+ requests per 10 seconds with low latency by leveraging batching, GPU time-slicing, and autoscaling policies.
5. Achieved 70%+ cost savings by migrating workloads from SageMaker Endpoints to Kubernetes (EKS), utilizing container reusability, spot GPU nodes, and resource sharing.

Education

BSc Computer Science in (Cloud Technology & Information Security)

Nagindas Khandwala College (Mumbai University), Mumbai | May 2021 - April 2024

Key Skills

- Cloud: AWS, Azure, GCP -
- CI/CD: Azure DevOps, GitHub Actions, Jenkins, GitLab, Code pipeline, etc. -
- Containers & Orchestration: Docker, Kubernetes,AKS, ECR, EKS, ACR. -
- Optimization: GPU time slicing, batching, quantization (familiar with concepts for LLM serving) -
- Security Tools: Trivy, OWASP Dependency Checker, OWASP ZAP, SonarQube, etc. -
- Monitoring: Azure Monitor, App Insights, CloudWatch, Prometheus, Grafana, Fluentd, ELK Stack, etc -
- Version Control: GitHub, Azure Repos, GitLab, etc. -
- Languages & Scripting: Python, Java, Bash, PowerShell, YAML -

Certifications

AWS Solutions Architect Associate (AWS) | September 2024

Microsoft Azure Fundamentals (AZ-900) (Microsoft Azure) | May 2022