Sabari Aditiyaa

Senior MLOps & LLMOps Engineer

sabari.ml | sabariaditiyaa@gmail.com | linkedin.com/in/sabariaditiyaa

Professional Summary

Senior MLOps & LLMOps Engineer with **5.6 years** of experience architecting scalable, production-grade AI platforms at **Bank of America** and **Infosys**. Expert in designing **Kubernetes-native inference infrastructure**, combining **KServe** and **vLLM** to deliver low-latency, high-throughput LLM serving. Proven track record of building robust **CI/CD pipelines** that reduced deployment lead time by **70%**. Specialized in **GPU resource optimization**, reliable data engineering for **50M+ records**, and ensuring **99.9% system uptime** for mission-critical enterprise workloads.

Technical Skills

Infrastructure & DevOps: Kubernetes (EKS/AKS), Docker, Terraform (IaC), Ansible, Jenkins, Git, Linux, CI/CD Pipelines, Helm, NVIDIA MIG

Cloud Platforms: AWS, Azure, GCP

ML Platforms & Services: AWS SageMaker, Azure Machine Learning, Azure AI Foundry

MLOps Orchestration: Kubeflow, MLflow, Airflow, ArgoCD (GitOps), **KServe**, **DVC**

LLM Inference & GenAI: vLLM, Ray Serve, KV Cache Optimization, Paged Attention, Prefill vs Decoding, Hugging Face, RAG

Observability & Monitoring: Prometheus, Grafana, ELK Stack, Dynatrace, **DCGM** (Data Center GPU Manager), CloudWatch

Professional Experience

Infosys Jul 2024 – Present

Associate Consultant – MLOps / LLMOps *Bengaluru, India*

- **LLM Platform Architecture:** Designed and operated a Kubernetes-native model-serving platform, supporting both traditional ML and Generative AI workloads.
- **Inference Orchestration (KServe):** Architected a serverless inference layer using **KServe** to orchestrate **vLLM** backends, enabling **scale-to-zero** capabilities to optimize GPU costs and implementing canary rollouts for risk-free production updates.
- **Performance Optimization:** Deployed **vLLM** with PagedAttention and **KV Cache optimization**, significantly improving token throughput and reducing latency under concurrent load.
- **Data & Artifact Management (DVC):** Implemented **DVC** within CI/CD pipelines to govern the lifecycle of large-scale model weights and datasets, ensuring full reproducibility across RAG ingestion workflows.
- **GPU Infrastructure:** Configured **NVIDIA MIG** (Multi-Instance GPU) and node affinity rules on Kubernetes to isolate workloads, maximizing hardware utilization for multi-model serving.
- **RAG Pipeline Ops:** Operationalized Retrieval-Augmented Generation (RAG) systems by building robust retrieval workflows that integrate seamlessly with standardized inference pipelines.
- **Training Orchestration:** Leveraged AWS SageMaker, Azure ML, and Azure AI Foundry for training orchestration, model registration, and deployment workflows.
- **Observability:** Implemented deep monitoring for inference workloads using Prometheus and Grafana, tracking key infrastructure metrics including GPU saturation, token latency, and throughput.

Bank of America Jul 2020 – Jul 2024

Software Engineer – Cloud DevOps / Infrastructure *Chennai, India*

- **Enterprise Infrastructure:** Engineered and operated secure, scalable cloud infrastructure across AWS, Azure, and GCP for mission-critical systems.
- **Security & Compliance:** Enforced strict banking-grade security policies including RBAC, VPC peering, and encryption at rest/transit, ensuring all infrastructure met rigorous audit standards.
- **Kubernetes Operations:** Managed production Kubernetes environments ensuring high availability, scalability, and automated workload scaling.
- **CI/CD Automation:** Architected and automated CI/CD pipelines using Jenkins, Git, Terraform, and Ansible, reducing deployment time and manual operational effort.
- **Infrastructure as Code:** Implemented Infrastructure as Code using Terraform, significantly reducing configuration drift and ensuring consistent environment provisioning.

- **Backend Engineering:** Built and supported high-throughput APIs using Java Spring Boot and Node.js on Kubernetes, ensuring robust backend performance under load.
- **Reliability Engineering:** Led automation initiatives that reduced system downtime by 25% through improved monitoring and self-healing infrastructure.
- **Monitoring Stack:** Implemented comprehensive monitoring and logging using CloudWatch, Prometheus, and ELK to ensure proactive system health tracking.

Key Projects

Enterprise AI Model-Serving & Inference Platform | *vLLM, KServe, Kubernetes*

- Built a standardized Kubernetes-based model-serving platform for ML and LLM inference services, decoupling application logic from model internals.
- Enabled LLM inference using **vLLM** with batching and cache-aware execution, optimizing resource utilization for high-demand workloads.
- Integrated **DVC** for model registry management, enabling seamless rollbacks and precise version control of heavy model artifacts.
- Implemented RAG workflows while ensuring modularity, allowing independent scaling of retrieval and generation components.
- Benchmarked latency, throughput, and resource utilization for continuous performance optimization.

Real-Time AI Voice Calling Platform (GenAI Infrastructure) | *Azure AI, Twilio*

- **Overview:** Designed and operated a real-time AI voice inference platform integrated with telephony systems for automated customer interaction.
- **Low-Latency Engineering:** Used Azure AI Foundry and Azure App Services to host low-latency inference endpoints, achieving **sub-200ms response latency**.
- **System Integration:** Integrated Twilio for inbound and outbound calling with real-time audio streaming, ensuring natural conversational flows.
- **Scalability:** Ensured scalability, availability, and observability under concurrent call load to support high-volume operations.

Verizon Ticket Management – Wireless Platform Enablement | *Spring Boot, CI/CD*

- **Scale:** Executed data migration pipelines transferring **50M+ records** with high accuracy, ensuring zero data loss during platform modernization.
- **Backend Engineering:** Built and operated CI/CD pipelines for middleware, backend, and UI deployments, streamlining the release process.
- **Infrastructure:** Managed infrastructure setup, scaling, and monitoring across environments to support enterprise wireless workflows.
- **Reliability:** Improved reliability using containerization and Kubernetes orchestration to handle enterprise-scale traffic.

Vehicle & Home Loan Management Platform | *AWS, Terraform, Docker*

- **Overview:** Designed and operated secure, scalable cloud infrastructure for a high-traffic loan platform handling sensitive financial data.
- **Infrastructure as Code:** Automated infrastructure provisioning using Terraform to ensure consistent and rapid environment scaling.
- **Release Automation:** Built CI/CD pipelines reducing release time by 40%, significantly accelerating feature delivery.
- **Containerization:** Deployed containerized services on Kubernetes ensuring high availability and fault tolerance.

Certifications & Education

AWS Certified Solutions Architect
AWS Certified Cloud Practitioner
Google Cloud Certified Associate Cloud Engineer
Oracle Cloud Infrastructure 2023 Foundations Associate
IBM Blockchain Foundation Developer & Docker Essentials

Bachelor of Technology in Information Technology
SRM Institute of Science and Technology, India (2016 – 2020)