Archit Dadhich

Chandigarh, India | dadhich.archit11@gmail.com | 9079183284

Introduction

Innovative DevOps Engineer with 1+ years of experience solving complex infrastructure challenges through automation, cloud-native solutions, and AI model deployment using Kubernetes, Terraform, and GitLab CI/CD. Passionate about optimizing workflows, enhancing observability, and delivering scalable, secure solutions

Experience

DevOps and Cloud Engineer, Xenonstack – Mohali, Punjab March 2024 - July 2025

- Configured a 4-node GPU cluster with Kubernetes, NVIDIA CUDA, and Kubeflow, deploying 10+ machine learning models and reducing inference latency by 25% for scalable, production-ready AI workflows.
- Built GitLab CI/CD pipelines with Docker and Helm, cutting deployment cycles by 35% for 5+ microservices, achieving zero-downtime releases.
- Automated AWS (EC2, EKS, S3) and Azure (AKS, Blob Storage) provisioning with Terraform, reducing setup time from 1 day to 2 hours and eliminating 90% of configuration errors.
- Configured Prometheus, Grafana, and Loki for real-time monitoring, slashing incident detection time by 40% and maintaining 99.95% application uptime.
- Implemented NVIDIA Triton Inference Server and MLflow for model serving and experiment tracking, improving data scientist efficiency by 20%.
- Partnered with data scientists and developers to integrate AI models into production, ensuring seamless deployments and 98% GPU resource utilization.
- Analyzed AWS resource usage with CloudWatch, implementing tagging strategies to reduce costs by 10% across 5+ services.
- Designed disaster recovery with Velero backups, reduced downtime risks 90%.
- Contributed significantly to documentation and audits, helping the team successfully achieve AWS DevOps and CloudOps Competencies

Projects

NexaStack AI Inference Cluster

- Designed an NVIDIA RTX 4090-based AI inference cluster with Kubernetes and Docker, deploying 20+ models with 99.5% uptime.
- Containerized 20+ LLMs (e.g., GPT variants) and SLMs (e.g., Phi-3) using CUDA and Kubeflow, achieving <50ms latency and 98% GPU utilization.
- Collaborated to integrate models via APIs, cutting integration time by 40% for 5+ use cases.
- Automated deployments with GitLab CI/CD and Helm, reducing deployment time to <30 minutes, with Prometheus/Grafana cutting incident response by 35%

Prominent US-based Dairy Manufacturing Workload Migration and Automation

- Migrated workloads from VMware to self-managed Kubernetes clusters with near-zero downtime, integrating seamless GitLab CI/CD pipelines aligned with DevSecOps principles.
- Enforced backup and restore mechanisms using Velero and MinIO for stateful and stateless workloads.
- Real-time observability was implemented with Prometheus and Grafana, integrating alert mechanisms for proactive responses together with the NOC and ITSM setup

Skills

Containerization & Orchestration:	Docker, Kubernetes, Helm, ArgoCD
CI/CD & IaC:	GitLab CI/CD, Terraform, Terragrunt, Jenkins(Basic)
Cloud Platforms:	AWS (VPC, IAM, EC2, EKS, S3), Azure (AKS, Blob Storage, Sentinel)
Monitoring & Observability:	Prometheus, Grafana, CloudWatch, Loki, SLI/SLO Tracking, Jaeger
AI Infrastructure:	GPU Cluster Management, AI Model Deployment (Inference and Training), MLflow
Scripting & Tools:	Bash, Python, Git, SonarQube, Trivy, Backstage, Kyverno
Practices:	GitOps, Disaster Recovery (Velero), FinOps, DevSecOps

Education

PG Diploma in Data Science, C-DAC, Bengaluru

B.Tech in Computer Engineering, Marwadi University, Rajkot

Certifications

Azure Administrator Associate