Mohsin Khan Shaik

AI Infrastructure - Site Reliability Engineer - Platform Engineer - Cloud Infrastructure Engineer
Michigan | +1 (248) 866-0447 | mohsinkhanshaik2@gmail.com | Linkedin

PROFESSIONAL SUMMARY

AI Infrastructure and Site Reliability Engineer with **3+ years of professional experience** designing and operating **large-scale distributed systems, cloud-native platforms, and Kubernetes-based AI/ML infrastructure**. Expertise in **AWS, Kubernetes, Terraform, Docker, and CI/CD**, with a strong focus on **scalability, reliability, and performance optimization**. Proven track record of improving **system availability, reducing latency, and automating infrastructure using Infrastructure as Code (IaC)**. Experienced in **GenAI infrastructure, model deployment pipelines, and platform engineering**, with hands-on experience in **observability, incident response, SLA/SLO management, and production-grade system design**. Skilled at building **high-availability, fault-tolerant systems** supporting mission-critical AI workloads.

TECHNICAL SKILLS

- **Programming & Scripting:** Python, Bash, SQL, Shell Scripting, YAML, JSON, REST APIs, Git
- **Cloud & Infrastructure:** AWS (EC2, S3, EKS, ECS, Lambda, IAM, VPC, CloudWatch), Kubernetes, Docker, Terraform, CloudFormation, Serverless Architecture, Infrastructure as Code (IaC), Networking Fundamentals (DNS, Load Balancing, TCP/IP)
- **CI/CD & DevOps:** Jenkins, GitHub Actions, GitLab CI/CD, Bitbucket Pipelines, CI/CD Pipeline Design, Release Engineering
- **Observability & Monitoring:** Prometheus, Grafana, Kibana, AWS CloudWatch, ELK Stack, Distributed Tracing (OpenTelemetry), Logging & Alerting Systems
- **Databases & Systems:** PostgreSQL, MySQL, DynamoDB, Redis, MongoDB, Linux Systems Administration, System Performance Tuning
- **SRE & Platform Engineering:** SLA/SLO Monitoring, Incident Management, Root Cause Analysis (RCA), High Availability, Disaster Recovery, Capacity Planning, Fault Tolerance, Reliability Engineering
- **AI Infrastructure & MLOps:** GenAI Infrastructure, AI Workloads, Model Deployment Support, AI Platform Automation, MLOps Fundamentals, Model Serving, Feature Store Concepts
- **Containerization & Orchestration:** Kubernetes (EKS), Helm, Container Orchestration, Auto-scaling, Service Mesh (Istio - basic knowledge)
- **Security & DevSecOps:** IAM Policies, Secrets Management, Infrastructure Security, CI/CD Security, Compliance & Policy Enforcement
- **Tools & Collaboration:** Jira, Confluence, Git, Agile, Scrum, On-call Operations

PROFESSIONAL EXPERIENCE

AI Infrastructure Reliability Engineer Jan 2025 – Present | USA
Coinbase

- Designed and operated **production-grade Kubernetes clusters** and cloud-native infrastructure on AWS using Terraform and Docker, improving **deployment reliability** and reducing provisioning time by **35%**
- Built and optimized **end-to-end CI/CD pipelines** using Python, Bash, and IaC, reducing **manual deployment effort by 45%** and improving release consistency
- Implemented **observability systems (metrics, logging, alerting)**, reducing **MTTR by 40%** and improving incident response efficiency
- Integrated **security controls and compliance checks** into CI/CD pipelines in collaboration with DevSecOps teams, reducing review delays by **30%**
- Managed **high-availability distributed systems**, ensuring **SLA/SLO adherence**, fault tolerance, and system resilience
- Led **incident response and root cause analysis**, reducing recurring incidents by **28%**
- Optimized **Kubernetes resource utilization**, improving performance and efficiency of AI infrastructure workloads

AI Infrastructure Engineer – GenAI & Cloud Optimization Feb 2022 – Dec 2023 | IND
CirrusLabs

- Built **automated infrastructure optimization frameworks** using AWS and Terraform, improving resource utilization across enterprise environments
- Developed **AI-driven infrastructure automation solutions** supporting GenAI workloads, reducing operational effort by **40%**
- Designed and maintained **Infrastructure as Code (IaC)** templates using Terraform and CloudFormation, reducing provisioning time by **50%**
- Collaborated with engineering teams to **optimize distributed systems performance** and enhance infrastructure efficiency
- Implemented **monitoring and observability pipelines** using AWS CloudWatch, improving system visibility and performance tracking
- Supported **containerized and serverless architectures**, improving application reliability and scalability

PROJECTS

Enterprise AI Infrastructure Monitoring & Automation Platform
AWS, Kubernetes, Terraform, Grafana, Python

- Designed a **cloud-native observability platform** for Kubernetes workloads using Grafana and centralized logging pipelines
- Automated **infrastructure provisioning workflows**, reducing configuration effort by **45%**
- Built **proactive alerting and auto-remediation systems**, reducing incident response time by **38%**
- Improved **cluster scalability and operational efficiency** for AI workloads

GenAI-Powered Infrastructure Optimization Platform
AWS, Python, Terraform, Lambda

- Developed a **GenAI-driven infrastructure optimization system** to identify inefficiencies and improve resource utilization
- Built automation workflows for **idle resource detection and compute optimization**, improving efficiency by **25%**
- Integrated **CloudWatch metrics and serverless automation**, enabling real-time monitoring and optimization
- Enhanced **infrastructure governance and observability** through automated dashboards

EDUCATION

Master's in Data Analytics | Indiana Wesleyan University Jan 2024 - Aug 2025 | USA
Bachelor's in Information Technology | Shadan College Of Engineering And Technology Jun 2018 - Jan 2023 | IND