

EHTISHAM MUBARIK

+1 (217) 200-0250 | ehtisham.m7@gmail.com | ehtishammubarik.com | linkedin.com/in/ehtisham-mubarik | github.com/ehtishammubarik

PROFESSIONAL SUMMARY

Senior AI Platform and DevOps Engineer with 7+ years in production AI and cloud infrastructure. I build and ship agentic systems: subagent orchestration, dynamic workflows, custom skill packaging, harness customization, and token-cost optimization for LLM and agent workloads. Recent work spans AWS Bedrock and event-driven ML for a real-time gaming platform, Ray-based distributed training and inference for AI-native product teams, and multi-cloud plus on-prem GPU orchestration on Kubernetes (MetalLB, Kamaji, Rancher). My graduate thesis on GPU Cluster Optimization for ML Workloads underpins deep expertise in GPU scheduling and cost efficiency. Track record: 60%+ AWS spend reduction, production SRE and content-generation agentic harnesses, and Pulumi/Go modules for a multi-tenant AI platform. CKA and CKAD certified.

CORE SKILLS

AI, MLOps & Agentic Systems: Kubeflow, MLflow, Ray, Airflow, PyTorch, LLM deployment, RAG, vector databases, Anthropic Claude Agent SDK, LangChain/LangGraph, subagent orchestration, dynamic multi-step workflows, custom skill packaging, agentic harness customization, token-cost optimization, Marvin, Modal.

Platforms & Infrastructure: AWS (Bedrock, Redshift, Glue, EKS), Azure, GCP, IBM Cloud, OpenStack (Nova, Neutron, Cinder, Keystone, Glance, Horizon, Ceilometer) with Ceph on bare metal, on-prem Kubernetes (MetalLB, K3s, Kamaji, Rancher), Gateway API, GPU orchestration, custom cluster provisioning, Pulumi, Terraform, ArgoCD, Ansible, Helm, VPC/CIDR/network design, OVS/VXLAN/DVR, Istio, VMware, Proxmox.

Reliability, Security & Compliance: Prometheus/Grafana/Loki, PagerDuty, SLA/SLO design, incident response, HIPAA, SOC 2, ISO 27001, IAM/RBAC, Vault/KMS, image hardening, air-gapped delivery, regression/load/stress and penetration testing.

Delivery & Leadership: GitHub Actions, GitLab CI, Jenkins, Argo Workflows, Kafka + event-driven pipelines, Python, Go, Bash, HCL, cross-functional team leadership, technical hiring, mentoring, cost optimization, stakeholder communication.

EXPERIENCE

Stack8s | Senior DevOps Engineer (Contract) Sep 2024 to **Present** | Remote — Kubernetes automation platform for AI/ML teams

- **Reduced cluster provisioning to under 10 minutes:** production-ready Kubernetes with GPU compute across AWS, Azure, and on-prem, via custom cluster provisioning with Kubeflow plug-and-play so teams leverage compute immediately.
- Built Kamaji-based multi-tenant Kubernetes with Gateway API for advanced routing and traffic management across tenant workloads.
- Ran Rancher-based cluster management across multi-cloud and on-prem environments, with GPU orchestration and intelligent workload distribution.
- Delivered vanilla Kubernetes on customer-owned hardware for data sovereignty in regulated AI and ML environments.
- Automated full OpenStack deployment on bare metal with Ceph as the storage backend, using custom Ansible roles, playbooks and group_vars (no containerised control plane): Nova, Neutron, Cinder, Keystone, Glance, Horizon and Ceilometer, with OVS/VXLAN/DVR networking, re-runnable for redeployment and cluster expansion.

Hyve Labs | Senior DevOps Engineer (Contract) Apr 2025 to Jul 2026 | Remote — real-time gaming platform

- Own end-to-end production of the gaming platform: real-time event fan-out, elastic multi-region autoscaling, low-latency multiplayer backends, and the full release lifecycle.
- **Delivered 60%+ reduction** of AWS spend through smarter model selection and inference-harness improvements on Bedrock, disciplined rightsizing of compute and storage, and senior operational best practices across the footprint.
- Rebuilt the AWS data and ML stack (Bedrock, Airflow, Redshift, Glue) around a Kafka-driven pub/sub architecture; event fan-out feeds analytics, model workflows, and downstream game systems at real-time scale.
- Shipped a Marvin-based agentic harness for game content generation and a LangGraph-driven SRE harness for incident triage, dynamic alert classification, and token-optimized remediation.
- Engineered a secure production layer (layered IAM/RBAC, secrets management, image hardening, network policy), ran regression/load/stress and periodic penetration testing, and owned PagerDuty escalation and P0 to P2 SLA enforcement.

Dressler Consulting | Sr DevOps & MLOps Consultant (Contract) Nov 2022 to Jul 2026 | Remote

- Designed Ray API and Ray Head cluster architectures on both on-prem (bare-metal Kubernetes + MetalLB) and cloud-managed environments for distributed AI model training and inference.
- Built Pulumi modules in Go for a multi-tenant AI platform: typed infrastructure code enabling per-tenant provisioning of Kubernetes, GPU pools, and MLOps stacks with automated teardown.

- Built an agentic harness for training-pipeline orchestration: subagent-based failure recovery, dynamic branching for hyperparameter sweeps, custom skill loaders per model family, and token-optimized calls for long-running jobs.
- Owned production alerting for distributed training (GPU utilization, cluster health, cost anomalies, silent-failure detection) across 3.5+ years of multiple client engagements.

[Aceso Analytics](#) | DevOps Engineer, Team Lead Apr 2020 to Sep 2024 | Norway

- Led the engineering team on a healthcare IoT platform serving LoRaWAN device fleets across multiple elderly-care facilities; architected the IoT and LoRaWAN stack on Kubernetes with HIPAA-aligned operational practices.
- Built scalable monitoring, alerting, and edge inference for real-time device fleets; mentored junior engineers and ran technical reviews.

Earlier DevOps roles — Afiniti, Halfpenny Software, Emumba | 2019 to 2021

- Automated build, packaging, and release of 40+ microservices across air-gapped enterprise environments; migrated legacy infrastructure to on-prem Kubernetes with production Ansible playbooks and Jenkins shared libraries.
- Ran parallel DevOps engagements for US and international clients on AWS/Azure/GCP (containerization and clustering migrations with Terraform, Ansible, Bash); built monitoring dashboards, CI/CD runners, and observability stacks for hybrid and on-prem customers.

CERTIFICATIONS

Certified Kubernetes Administrator (CKA) and Certified Kubernetes Application Developer (CKAD), The Linux Foundation.

EDUCATION

MSCS, National University of Sciences and Technology (NUST), Islamabad, 2019 to 2021. *Thesis: GPU Cluster Optimization for ML Workloads.*

BSCS, National University of Computer and Emerging Sciences (FAST-NU), 2015 to 2019.

Community: Organize internal AI/MLOps/agentic-tooling workshops for engineering teams, attend cloud and infrastructure meetups, and publish on production AI and agentic system operations.