Deepak Inugala

Senior DevOps Engineer | MLOps & LLMOps | Agentic AI Infrastructure | International Client Delivery

Email: deepak.1990@hotmail.com **Phone:** +971 50 494 5921

Location: Abu Dhabi, UAE **LinkedIn:** linkedin.com/in/deepak-inugala

10+ yrs	6 Countries	3 Verticals	GPU Cluster	40% Travel	99.99%
Total Experience	International Delivery	GOVINT/OSINT/Smart Nation	Management Experience	Quarterly On-Site	Platform Uptime

PROFESSIONAL SUMMARY

Senior DevOps / SRE Engineer with 10+ years of experience architecting cloud-native, on-premises, and air-gapped infrastructure for enterprise and government clients. Specialist in MLOps, LLMOps, and Agentic AI with hands-on GPU cluster management, vLLM production serving, and autonomous AI platforms. Proven international delivery engineer deployed G42 platforms on-site across Kazakhstan, Angola, Bahrain, and UAE, and remotely for Maldives and Ethiopia, engaging clients across GOVINT, OSINT, and Smart Nation verticals. Expert across the full DevOps lifecycle: IaC, CI/CD, GitOps, Kubernetes, observability, and security.

TECHNICAL SKILLS

AI / LLM & Agentic	vLLM, Triton, Ray Serve, Ollama, OpenAI-compatible APIs, LangChain, LangGraph, HuggingFace Hub, ONNX Runtime, LiteLLM, OpenClaw, NemoClaw, n8n, RAG (Qdrant), MCP Servers
GPU & MLOps	NVIDIA GPU Clusters (H100/H200/A100), MIG Config, DCGM Exporter, GPU Operator, NCCL, InfiniBand; MLflow, Kubeflow Pipelines, Argo Workflows, Seldon Core, BentoML, DVC, W&B
Cloud Platforms	Azure (AKS, Azure ML, Key Vault, Defender, Sentinel, Monitor), AWS (EKS, EC2, S3, RDS, Lambda, CloudWatch), G42 Cloud, Huawei Cloud, OpenStack
Kubernetes & IaC	AKS, EKS, Rancher, RKE2, Helm, ArgoCD, Kustomize, Docker, Containerd, Harbor; Terraform, Ansible, GitLab CI/CD, GitOps, Python, Bash
Observability & Sec	Prometheus, Grafana, ELK Stack, DCGM, Zabbix, Loki; Palo Alto Firewalls, Sentinel SIEM, Defender, Zero-Trust, Keycloak OIDC, Falco, Trivy
Databases & Tools	ClickHouse, PostgreSQL, MySQL, Oracle DB, Elasticsearch, Confluent Kafka, Ubuntu, RHEL, CentOS, Windows Server; Ceph, SAN, NAS, Airflow

WORK EXPERIENCE

Senior DevOps Engineer | SRE | MLOps / LLMOps | Group 42 (G42)

07/2019 – Present | Abu Dhabi, UAE | On-Site: Kazakhstan, Angola, Bahrain | Remote: Maldives, Ethiopia

International Delivery — On-Site & Remote (6 Countries)

- Deployed and commissioned G42 platforms (GOVINT, OSINT, Smart Nation) at client sites across Kazakhstan, Angola, and Bahrain acting as sole technical authority for infrastructure, SRE, DevOps, and security; additionally supported Maldives and Ethiopia remotely via secure VPN, coordinating with local client IT teams across time zones.
- Managed all pre-deployment readiness: data center vendor coordination, rack-and-stack, power/cooling validation, network setup, hardware commissioning, and go-live UAT travelling approximately 40% of each quarerly across international sites.
- Led client technical onboarding and structured training programs at each site facilitating knowledge-transfer workshops, producing localized runbooks and handing over fully operationalized systems with SLA documentation and support escalation paths.
- Concurrently managed infrastructure and SRE across GOVINT (secure air-gapped government intelligence platforms), OSINT (large-scale public data ingestion via Kafka/ELK with GPU-accelerated NLP) and Smart Nation projects (IoT aggregation, city-scale analytics, AI inference stacks and government ministry API gateways).

AI Infrastructure, MLOps & LLMOps

- Designed and provisioned bare-metal enterprise GPU clusters; configured MIG profiles for concurrent LLM inference and training; deployed NVIDIA GPU Operator with DCGM Exporter integrated into Prometheus/Grafana — reducing GPU incident MTTR by 40%.

- Built production vLLM serving stacks for large-scale LLMs (7B-72B) with LiteLLM gateway, Open WebUI, and Mattermost; architected end-to-end MLOps pipelines using Kubeflow Pipelines and Argo Workflows — reducing model release cycles from weeks to hours.
- Deployed RAG pipelines over SRE runbooks using Qdrant and LangGraph; built n8n workflow automation reducing operational toil by 50%, configured HuggingFace Hub mirroring in air-gapped environments with GPTQ/AWQ quantization, implemented canary and blue-green model rollouts via Argo CD.

Platform Reliability & Cloud Infrastructure

- Architected and maintained Kubernetes clusters (AKS, EKS, Rancher, RKE2, bare metal) supporting 50+ production AI/ML services achieving 99.99% uptime; automated provisioning with Terraform and Ansible across Azure, AWS, and G42 cloud reducing environment build time from days to under 2 hours.
- Designed Azure security architecture aligned to ISO 27001 and UAE IA standards (Key Vault, Sentinel SIEM, Defender for Cloud, Palo Alto firewall, Zero-Trust) built full-stack observability across 200+ nodes (Prometheus/Grafana, ELK Stack, Zabbix, DCGM) led incident response achieving 60% reduction in P1 incidents.

Cloud Engineer | First Abu Dhabi Bank (FAB)

02/2018 – 07/2019 | Abu Dhabi, UAE

- Cloud Administration & Infrastructure: completed FGB-NBAD bank merger IT integration with zero major incidents; designed and managed highly available AWS architectures (EC2, EKS, RDS, S3, Lambda, ECS, Load Balancers) with Well-Architected Framework reviews, centralized monitoring via CloudWatch, Prometheus, and Grafana.
- Elasticsearch & Hadoop Administration: deployed and managed multi-node Elasticsearch clusters (ILM policies, shard allocation, Kibana dashboards); installed and administered Cloudera CDH clusters (HDFS NameNode HA, YARN, Hive, HBase, Spark, Zookeeper) with Kerberos and Ranger authorization.
- Vulnerability Patching: owned full vulnerability management lifecycle — triaged CVEs by severity, applied OS and middleware patches via Ansible, validated remediation through post-patch scanning; led incident response reducing repeat incidents by 35%.

Linux Engineer | HCL InfoSystems Limited

04/2015 – 01/2018 | Abu Dhabi, UAE

- Core Linux Administration & Data Center Operations (3,000+ Physical Servers): managed large-scale Linux fleet (RHEL, CentOS) OS provisioning, Ansible configuration management, user administration, performance tuning, hardware fault diagnosis (iDRAC/iLO), and server firmware lifecycle across Dell, HP, and IBM servers.
- Network Switch & Firewall Administration: configured and managed Cisco and IBM switches (VLAN creation/management, trunk/access ports, STP/RSTP, LACP port-channels); administered firewall rule sets and access control policies — implemented DMZ segmentation, NAT rules, and ingress/egress controls.
- On-Premises-to-Cloud Migration: led strategic migration of Linux infrastructure and Cloudera Hadoop clusters to cloud with under 2 hours downtime and 0% data loss; managed full project delivery including timeline, budget, and stakeholder communication for 20+ node cluster.

CERTIFICATIONS

Microsoft Certified: Azure Administrator Associate (AZ-104)
Valid: 02/2026 - 07/2027

Microsoft Certified: Azure Solutions Architect Expert (AZ-305)
Valid: 07/2025 - 07/2027

AWS Certified Solutions Architect - Associate
Valid: 07/2022 - 07/2025 | ID: PFR7HTQ25EFQQS9T

Certified Kubernetes Administrator (CKA)
Valid: 09/2021 - 09/2023 | Renewal in Progress

G42 Cloud Certified Engineer
Valid: 10/2021 - 12/2024 | ID: G42C/SVD/CRT/0475

Red Hat Certified Engineer (RHCE)
Valid: 02/2015 - 02/2018 | ID: 150-012-904

EDUCATION

Master of Business Administration (MBA)
Jawaharlal Nehru Technology University | 2012 - 2015

Bachelor of Technology (B.Tech)
Jawaharlal Nehru Technology University | 2008 - 2012

LANGUAGES & SOFT SKILLS

Languages
English & Hindi (Professional) | Telugu (Native) | German (Elementary)

Soft Skills
Cross-functional Collaboration | Stakeholder Communication | Problem Solving | Rapid Learning | Adaptability