Deepak Inugala

Senior Site Reliability Engineer | AI Platform SRE | Cloud Infrastructure | International Delivery

Email: deepak.1990@hotmail.com
Phone: +971 50 494 5921
Location: Abu Dhabi, UAE
LinkedIn: linkedin.com/in/deepak-inugala

10+ yrs	60% Less	200+ Nodes	6 Countries	Verticals	99.99%
SRE / DevOps Experience	P1 Incidents Achieved	Observability Coverage	International Delivery	GOVINT/OSINT/Smart Nation/RND	Uptime Delivered

PROFESSIONAL SUMMARY

Senior Site Reliability Engineer with 10+ years of experience designing and operating production infrastructure at enterprise and government scale. Expert in SLO/SLI/error budget frameworks, incident management, full-stack observability, and reliability-first automation. Hands-on experience with enterprise GPU clusters, vLLM serving, Kubernetes (AKS, RKE2, bare metal), Terraform, Prometheus/Grafana/ELK, and Azure security architecture. Deployed and operated G42 reliability-critical platforms on-site in Kazakhstan, Angola, Bahrain, and UAE and remotely for Maldives and Ethiopia — defining SLOs, establishing monitoring, training client teams, and handing over fully operationalized systems across GOVINT, OSINT, and Smart Nation verticals.

TECHNICAL SKILLS

SRE & Reliability	SLO/SLI/Error Budget Design, Incident Management, Blameless Post-Mortems, Capacity Planning, Disaster Recovery, Chaos Engineering (LitmusChaos), Toil Reduction
Observability	Prometheus Operator, Grafana, Loki, Alertmanager, ELK Stack (ES/Logstash/Kibana/Filebeat), DCGM GPU Metrics, Zabbix, Azure Monitor, Fluent Bit, Sloth SLO
Cloud & Kubernetes	Azure (AKS, Key Vault, Defender, Sentinel, Monitor), AWS (EKS, EC2, S3, RDS, Lambda, CloudWatch), G42 Cloud; AKS, EKS, Rancher, RKE2, Helm, ArgoCD, Docker, Harbor
AI / GPU SRE	Enterprise GPU Cluster Management, MIG Partitioning, DCGM Exporter, vLLM, Kubeflow Pipelines, Argo Workflows, MLflow, LangGraph, RAG (Qdrant), n8n
IaC & Security	Terraform, Ansible, GitLab CI/CD, GitOps, Python, Bash; Palo Alto Firewalls, Sentinel SIEM, Defender, Zero-Trust, Keycloak OIDC, Falco, Trivy, OPA Gatekeeper
Databases & OS	ClickHouse, PostgreSQL, MySQL, Oracle DB, Elasticsearch, Confluent Kafka, Airflow, Ubuntu 22.04/24.04, RHEL 8/9, kernel tuning, containerd, CIS hardening; Ceph, SAN, NAS

WORK EXPERIENCE

Senior Site Reliability Engineer | 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 50% of each month 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).
- Managed enterprise network infrastructure including VLAN and IP address management (IPAM), Palo Alto and Azure Firewall administration, and Site-to-Site (S2S) and Point-to-Site (P2S) VPN configurations to ensure secure, scalable, and highly available hybrid network environments

SLO / SLI / Error Budgets & Incident Management

- Designed SLO/SLI frameworks across 50+ production services (availability 99.9-99.99%, latency P95/P99, error-rate targets); established error budget enforcement policies and introduced burn-rate alerting via Sloth (Prometheus SLO generator) replacing threshold-based alerting.
- Owned full incident management lifecycle reducing P1 incidents by 60% over 2 years, reduced MTTD from 15 minutes to under 3 minutes; eliminated 70% of false-positive pages; led blameless post-mortems with mandatory Jira-tracked action items; facilitated annual DR testing achieving RTO under 2 hours and RPO under 15 minutes.

Full-Stack Observability & AI Platform SRE

- Deployed kube-Prometheus-stack across all clusters with custom scrape configs for GPU metrics, Longhorn, etcd, and all workloads; built 30+ Grafana dashboards covering cluster health, GPU utilization/temperature, storage IOPS, network throughput, and application golden signals; deployed Loki + Fluent Bit for correlated metrics-and-logs debugging.
- Managed enterprise GPU node pools on AKS and RKE2; built vLLM serving stacks with DCGM-based HPA and P99 TTFT under 200ms at 1,000+ concurrent requests; deployed LangGraph SRE agent workflows reducing overnight on-call interventions by 40%.

Automation, IaC & Cloud Security

- Automated cloud provisioning with Terraform and Ansible across Azure, AWS, and G42 sovereign cloud reducing environment build time from days to under 2 hours; achieved 50%+ toil reduction via Ansible playbooks and n8n workflow automation, deployed LitmusChaos chaos engineering to validate resilience before production changes.
- 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.

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 Name Node HA, YARN, HBase, Hive, Spark, Zookeeper) with Kerberos and Ranger authorization, Vulnerability Patching

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.
- 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