RAJ SASIDHARAN

South San Francisco, CA | rajsasc@gmail.com | 609-815-6287

PROFESSIONAL SUMMARY

Senior DevOps / Platform Engineer with 23 years running secure and compliant cloud operations across AWS and hybrid environments, most recently in HIPAA-regulated healthcare and genomics at Myriad Genetics. Hands-on across cloud infrastructure, DevSecOps, CI/CD automation, Infrastructure as Code with Terraform and CloudFormation, and platform reliability. Built zero-trust security end to end: default-deny network policy, WireGuard encryption, and a Sigstore-signed supply chain verified at admission, with CIS-aligned hardening and SLO-based monitoring behind it. I take on the stalled and high-risk work, and I mentor the engineers around me.

KEY HIGHLIGHTS

Built end-to-end **LLM fine-tuning pipeline** on EKS: Axolotl-based QLoRA on Llama 3.1 8B running as a Kubeflow PyTorchJob, S3-backed adapter checkpoints via Pod Identity, served back through **vLLM --enable-lora** and closed the loop with parallel **base-vs-LoRA evaluation** via lm-eval against in-cluster vLLM endpoints.

Productionized **multi-model LLM serving with vLLM**, 9 models including Llama 3.1 8B/405B AWQ, Llama 3.2 11B Vision, Mixtral 8x7B AWQ, DeepSeek-R1-Distill-Llama-70B, and Llama Guard 3 8B for safety; 4-tier classifier in LangGraph (fast / reasoning / safety / tuned-LoRA) with Microsoft Presidio for PII redaction.

Implemented **scale-to-zero LLM serving via KEDA** (composite Cron + Prometheus vllm:num_requests_waiting triggers) layered over **Karpenter GPU autoscaling** across 5 instance types and Mountpoint S3 CSI for direct model-weight access, which removed idle GPU cost without hurting latency.

Drove DevSecOps hardening: Sigstore cosign + Kyverno keyless signing and admission-time signature verification, default-deny CiliumNetworkPolicies + WireGuard pod + node encryption for east-west zero-trust, Istio Gateway API for north-south TLS termination, and HA HashiCorp Vault demonstrating both Agent Injector and Vault Secrets Operator patterns.

Migrated apps to **Argo Rollouts** canary deployments with Prometheus-based AnalysisTemplates for automated promote/rollback, Istio + Gateway API traffic shifting, and oauth2-proxy + Keycloak OIDC authenticating the dashboard.

Instrumented full **LLM observability** with Langfuse v3 (ClickHouse-backed) for prompt/completion/cost tracing, OpenTelemetry + Grafana Tempo for distributed traces, Prometheus + Grafana with SLO multi-window multi-burn-rate alerts, and DCGM exporter for live GPU telemetry.

Re-architected the production EKS data plane, migrating from AWS VPC CNI + EKS Managed Node Groups to Cilium eBPF CNI + Karpenter, with a single-workload Fargate bootstrap (only the Karpenter controller itself) and 100% of remaining capacity, including CoreDNS, provisioned on demand by Karpenter on EC2. Cilium runs with full kube-proxy replacement, ENI IPAM, WireGuard pod + node encryption, and Gateway API. Added Hubble L7 observability and Tetragon kernel-level runtime threat detection.

CORE SKILLS

Cloud & Platform

AWS (EC2, S3, RDS, VPC, IAM, KMS, Route 53, CloudFormation, Pod Identity), EKS, OpenShift/ROSA, Kubernetes, Karpenter, Docker, Podman

CI/CD & GitOps

GitHub Actions, Argo CD (App-of-Apps + ApplicationSets), Argo Rollouts, Kustomize, Helm, GitOps, GitHub Enterprise Server, Jenkins, GoCD

LLM Serving / Fine-Tuning

vLLM (multi-model + LoRA), Axolotl QLoRA, Kubeflow Training Operator (PyTorchJob), AWQ quantization, Llama 3.1 / 3.2 / 405B, Mixtral 8x7B, DeepSeek-R1, Llama Guard 3

AI Platform

RAG pipelines, LangGraph agentic orchestration, Qdrant vector DB, bge-m3 embeddings + bge-reranker, Langfuse (ClickHouse), ragas-eval, lm-eval, Microsoft Presidio (PII)

Autoscaling & Resilience

KEDA (Cron + Prometheus triggers), prometheus-adapter, HPA on custom metrics, Karpenter GPU NodePools, Argo Rollouts canary + AnalysisTemplate

Service Mesh & Networking

Istio (Gateway API ingress, north-south only), Cilium eBPF (ENI IPAM, kube-proxy replacement, CiliumNetworkPolicy, WireGuard pod + node encryption, Gateway API), Hubble L7, Tetragon, AWS Load Balancer Controller, External-DNS, Gateway API CRDs

Identity & Security

HashiCorp Vault (HA + Agent Injector + VSO), Keycloak OIDC, OAuth 2.0/OIDC, AWS KMS, External Secrets Operator, cert-manager + ACME, Sigstore cosign + Kyverno, SAML/SSO, Active Directory, LDAP, PKI/CA, Zero Trust architecture, Secure SDLC, CIS-aligned hardening, HIPAA-regulated operations

Observability

Prometheus, Grafana, OpenTelemetry, Grafana Tempo, Langfuse, DCGM, Splunk, Datadog, Sentry, CloudWatch, SLO/error-budget alerting, incident response

Automation & Config

Terraform, Puppet, Ansible, Bash, Python, HCL

Systems & Data

Linux (RHEL, AL2023, Alpine), VMware ESX/vSphere, DNS/BIND, PostgreSQL, ClickHouse, Apache Kafka (Strimzi), Redis, MinIO/S3

PROFESSIONAL EXPERIENCE

Myriad Genetics | California

Senior DevOps Engineer / SRE IV **Jul 2017 – Jul 2026**

Architected the foundational EKS production-pattern platform with Terraform (cluster, VPC integration, IAM, ECR, Karpenter NodePools) and an ArgoCD App-of-Apps GitOps delivery model, bootstrapping HashiCorp Vault, External Secrets Operator + Vault Secrets Operator, cert-manager with Let's Encrypt-backed wildcard TLS, Cilium eBPF CNI, Istio (Gateway API only), Karpenter, and Kyverno via initial kubectl apply, then handing all ongoing lifecycle to ArgoCD for fully GitOps-driven delivery of every downstream workload.

Migrated the production EKS data plane from AWS VPC CNI and EKS Managed Node Groups to Cilium eBPF CNI with Karpenter. Only the Karpenter controller itself runs on a Fargate bootstrap profile (the single workload that must exist before any EC2 node can be provisioned); 100% of remaining capacity, CoreDNS (initially on Fargate, moved to EC2 after Fargate cold-starts during Karpenter consolidation caused cluster-wide DNS outages), Cilium operator, cilium-agent DaemonSet, Hubble Relay/UI, ArgoCD, EBS CSI, Istio, Tetragon, Prometheus stack, KEDA, and every application workload, runs on Karpenter-provisioned EC2 NodePools pinned via karpenter.sh/nodepool selectors. Workloads use Cilium ENI IPAM, full kube-proxy replacement, Gateway API, and WireGuard pod + node encryption. Authored 36 CiliumNetworkPolicies (with correct cross-namespace egress via toEntities: cluster), and added 7 Tetragon TracingPolicies for kernel-level runtime threat detection alongside Hubble L7 observability. This removed the AWS VPC CNI's per-node ENI-exhaustion ceiling and providing a unified eBPF data plane for both networking and security.

Built end-to-end **LLM fine-tuning pipeline** on EKS, **Kubeflow Training Operator** orchestrating PyTorchJobs running an Axolotl QLoRA config on Llama 3.1 8B (bf16, flash-attention, gradient checkpointing on a single g6.xlarge A10G GPU); trained adapters synced to S3 via Pod Identity, served back through **vLLM --enable-lora**, and closed with **parallel base-vs-LoRA evaluation** via lm-eval (winogrande, hellaswag) against in-cluster vLLM endpoints with results persisted to PostgreSQL for prompt/retrieval tuning feedback loops.

Productionized a **multi-model LLM serving fleet via vLLM** (9 models): Llama 3.1 8B for fast inference, **Llama 3.1 405B AWQ** for high-end reasoning, Llama 3.2 11B Vision for multimodal, **Mixtral 8x7B AWQ**, **DeepSeek-R1-Distill-Llama-70B**, plus bge-m3 embeddings and bge-reranker (cross-encoder), and **Llama Guard 3 8B** as the safety tier; model weights staged to S3 and accessed via the **Mountpoint S3 CSI driver**, eliminating per-pod download cost on cold start. LangGraph router exposes a **4-tier classifier** (fast / reasoning / safety / tuned-LoRA) with cost/latency-aware model routing and request-payload override for A/B testing.

Implemented **three-tier autoscaling** for AI workloads: **KEDA scale-to-zero** for vLLM with a composite Cron-schedule (business-hours floor) + Prometheus queue-depth (vllm:num_requests_waiting) trigger; **Karpenter GPU provisioning** across 5 instance types (g4dn, g5.12, g6, g6e, g6.12) with vLLM image pre-pull on node join to cut cold-start; **prometheus-adapter** to expose custom metrics for HPA on non-KEDA Deployments, HPA behaviors tuned for responsive scale-up (100% per 30s) and conservative scale-down (10-min stabilization) to prevent flapping.

Migrated four production services (chat-ui, ingestion-service, langgraph-service, rag-service) to **Argo Rollouts canary deployments** with **Prometheus-based AnalysisTemplates** for automated promote/rollback; canary traffic shifting routed via **Istio + Gateway API** HTTPRoute weights, with **oauth2-proxy + Keycloak OIDC** authenticating the Rollouts dashboard, and Argo CD ignoreDifferences configured for canary-mutated fields.

Implemented east-west zero-trust security with Cilium: default-deny CiliumNetworkPolicies across all namespaces and WireGuard pod + node encryption (replacing an earlier Istio sidecar mesh, which had mesh-wide STRICT mTLS PeerAuthentication + 33+ AuthorizationPolicy allow rules, after the Cilium migration moved east-west enforcement into the eBPF data plane). Retained Istio for north-south Gateway-API ingress only (chosen over Cilium Gateway API after a persistent cilium#45871 eBPF redirect bug), with a shared Gateway + per-app HTTPRoutes, automated TLS via cert-manager, and Route53 records via External-DNS, fully decommissioned legacy ingress-nginx in a 13-phase migration.

Hardened the supply chain end-to-end with **Sigstore cosign + Kyverno**, **keyless signing via GitHub Actions OIDC** trust policies (separate IAM role per app), and 6 Kyverno ClusterPolicies enforcing image-registry allowlist + cosign signature verification + **digest pinning at admission** for chat-ui, rag-service, langgraph-service, ingestion-service, training, and ragas-eval images, plus upstream verification of ArgoCD images.

Deployed **HashiCorp Vault HA** (Raft, 3 replicas) demonstrating both delivery patterns: **Vault Agent Injector** sidecars (per-pod identity, file-based templated secrets) for rag-service, and **Vault Secrets Operator (VSO)** for native K8s Secret sync into ArgoCD, Grafana, and Keycloak; **dynamic Postgres credentials** via Vault's database engine; **migrated 8 IRSA roles to EKS Pod Identity** for cleaner cross-account-free authentication.

Instrumented full **LLM observability stack**, **Langfuse v3 (ClickHouse-backed)** for prompt/response/token/cost tracing, **OpenTelemetry + Grafana Tempo** for cross-service distributed traces (RAG gateway → embedder → vector search → LLM), Prometheus + Grafana dashboards for p50/p95/p99 latency and model-routing splits, **SLO multi-window multi-burn-rate alerts** (Google SRE methodology), severity-tiered Alertmanager receivers, and DCGM exporter + NVIDIA dashboards for live GPU telemetry; added **Microsoft Presidio** (analyzer + anonymizer) for PII detection upstream of the ingestion pipeline.

Built **LangGraph agentic orchestrator** with a **think → act → observe loop** and four tools (RAG search, inventory lookup, Neo4j graph query, final answer); enforced **production-grade guardrails** (max-step limit, token budget, Redis kill switch); every reasoning step, tool call, and cost event publishes to dedicated Kafka topics (agent.trace, agent.action, agent.cost), enabling session replay, horizontal tool scaling via consumer groups, and durable audit independent of pod lifecycle.

Provisioned a **multi-tenant data layer** as an ArgoCD-managed Application: **Kafka (Strimzi, KRaft mode, TLS 1.3** with per-service X.509 client certs and topic-scoped ACLs), **Qdrant vector database**, Neo4j Enterprise with Keycloak

OIDC SSO, PostgreSQL, Redis, and MinIO, all packaged as Kustomize manifests with credentials synced from Vault via External Secrets Operator; authored custom Lua health checks so ArgoCD could track Strimzi CRD readiness, and enforced namespace isolation with **default-deny NetworkPolicies**.

Led migration of **counsyl.com DNS domain** from Linux BIND to Active Directory–integrated DNS, partnering with the Windows team to deliver a more maintainable and secure name resolution architecture; owned ongoing **platform operations** including SSL/TLS certificate lifecycle management, container image upgrades, and security patching across ROSA + EKS environments, ensuring uninterrupted availability for patient-facing services.

Brought in to stabilize a fragile infrastructure environment; led foundational platform and systems modernization efforts that directly improved reliability for laboratory robotics and genomic processing workflows.

Designed and deployed new VMware ESX virtualization infrastructure with shared storage and bonded networking, replacing fragile local-storage and Vagrant-based lab environments and eliminating recurring VM failures.

Took ownership of a stalled identity modernization initiative: built Active Directory infrastructure from scratch, migrated 500+ user accounts from OpenLDAP, and integrated OneLogin/SAML for single sign-on across the organization.

Partnered with network engineering to launch secure Wi-Fi infrastructure, implementing an internal Windows Server 2016 Certificate Authority for client certificate issuance and strengthening wireless security posture.

Supported early containerization efforts, helping migrate services from EC2-based deployments toward Amazon ECS and later OpenShift, laying the groundwork for the GitOps-based platform adopted post-acquisition.

Autodesk Inc. (via Right Click Systems LLC) | California

DevOps Engineer / SRE **Aug 2014 – Jul 2017**

Architected and deployed AWS infrastructure from the ground up, including VPC design, CloudFormation templates, EC2/EBS provisioning, IAM policies, Route 53, CloudFront, and CloudWatch monitoring.

Designed centralized authentication for Unix hosts using Active Directory, OpenLDAP, and SSSD, unifying identity management across Red Hat Linux environments and eliminating per-host local account sprawl.

Built Python automation to generate Puppet and Nagios configurations, reducing manual setup time and supporting production operations across Docker, Kubernetes, and GoCD-based CI/CD environments.

ADDITIONAL EXPERIENCE

Tegile.com, Newark, CA, Solutions Engineer (Feb 2013 – Mar 2014)

Accenture Inc, Team Lead (May 2012 – Feb 2013)

EVault Inc, Emeryville, CA, Sr. Systems Administrator (Jan 2009 – Mar 2012)

Walmart.com, Brisbane, CA, Sr. Systems Administrator (Jun 2008 – Jan 2009)

Wellsfargo.com, San Francisco, CA, Systems Administrator (Feb 2008 – Jun 2008)

Earlier roles: Hewlett Packard, IPsoft, Tata InfoTech ATC (2003–2007), Systems & Unix Administration

CERTIFICATIONS

AWS Certified Solutions Architect

Red Hat Certified Engineer (RHCE)

Sun Certified System Administrator

Sun Certified Network Administrator

EDUCATION

B.E. in Electronics and Communication Engineering, Madras University, India