SUDHIR JAISWAL

Mumbai, India | +91-9839774797 | jaiswalsudhir2944@gmail.com
linkedin.com/in/sudhir-jaiswal-01a74836a

SUMMARY

Senior Cloud & AI Engineer building production platform and GenAI infrastructure on GCP and Kubernetes across media, BFSI, and telecom. Led the AWS-to-GCP migration of 100+ microservices, built AI-assisted root-cause analysis that cut MTTR by 35%, and ship multimodal AI pipelines (1 GB/min media streams), RAG, and LLM-evaluation systems.

EXPERIENCE

Searce Inc. Apr 2024 - Present
Senior Cloud & AI Engineer Mumbai, India
Promoted Jul 2026. Previously Cloud & AI Engineer, Apr 2024 - Jun 2026.

- Built an **AI-assisted root-cause analysis system** correlating GCP and AWS telemetry, reducing MTTR by **35%**. Instrumented services with OpenTelemetry collectors feeding Loki and Grafana for metrics, traces, and logs.
- Built AI automation for a live news-media pipeline processing streaming media at **1 GB/min**, using multimodal models to cluster and group related content and LLMs to classify and prioritize assets.
- Automated **BFSI banking-document workflows** with multi-process pipelines per document type (cheques, forms, policies). Masked sensitive content with **Data Loss Prevention** models before processing, then embedded it with a multilingual embedding model into a low-latency vector search index.
- Wrote **LLM evaluations across multiple models and modalities** and integrated them into CI/CD as a pre-release quality gate.
- Built a self-service platform for production-ready GKE environments, cutting developer onboarding time by **60%**. Wrote a Go controller for namespaces, quotas, and Helm releases, cutting manual provisioning by **70%**.
- Led migration of **100+ microservices from AWS to GCP**, moving Lambda functions to GKE. Built custom **Kong Gateway** plugins to keep downstream services backward compatible throughout the cutover.
- Automated infrastructure with **Terraform and GitLab CI**, building reusable modules for GKE, networking, and IAM. Standardized CI/CD across **50+ microservices** with ArgoCD ApplicationSets for multi-cluster sync and drift detection.
- Delivered a **zero-downtime Apigee Edge-to-X cutover**, migrating **120+ API proxies** with policies, KVMs, and apps. Built GCP service extensions for auth and rate limiting; authored design docs and delivery templates adopted across teams.

Madeline & Co. Jan 2024 - Apr 2024
Software Engineer Intern Remote

- Cut initial load times by **30%** with Next.js code-splitting. Added Jest tests and lint gates to CI.
- Built responsive Next.js and React front-end features through code reviews and iterative releases.

PROJECTS

KubeDoc AI *Multi-agent RAG platform*

- Built a FastAPI gateway with JWT authentication and rate limiting, LangGraph orchestration, pgvector hybrid search, Redis semantic caching, and Celery ingestion. Delivered cited answers over SSE; deployed on Kubernetes with Helm and HPA.

CognitoShield *Corrective RAG with PII guardrails*

- Tokenized PII at ingestion to keep sensitive values out of the vector index. Added bounded agent iterations, pluggable storage backends, and **24 offline tests**. Earned Honourable Mention at the Google Cloud Agent Builders Challenge (APAC 2026).

TECHNICAL SKILLS

AI & GenAI: LLMs, Multimodal Models, RAG, LLM Evaluations, DeepEval, Prompt Engineering, LangGraph, Google ADK, Vertex AI, Embeddings, Vector Search, pgvector, Data Loss Prevention (DLP), PII Guardrails
Cloud & Platform: GCP, AWS, GKE, Kubernetes, Docker, Helm, Microservices, AWS Lambda, Linux
DevOps & IaC: Terraform, Infrastructure as Code, GitOps, ArgoCD, CI/CD, GitLab CI, GitHub Actions
Backend & APIs: Python, Go, Bash, SQL, FastAPI, REST APIs, API Gateway, Kong, Apigee, PostgreSQL, Redis, Celery, Next.js, React
Observability: OpenTelemetry, Prometheus, Grafana, Loki, Monitoring & Alerting

EDUCATION

B.Tech, Computer Science and Engineering, IIIT Vadodara 2024

CERTIFICATIONS

Google Cloud: Professional Cloud Architect | Associate Cloud Engineer