

Sai Pisey

Senior DevOps Engineer | Site Reliability Engineer (SRE)

piseysai02@gmail.com | +91 93723 41689 | linkedin.com/in/sai-pisey | Mumbai, India (Remote)

SUMMARY

DevOps Engineer and Site Reliability Engineer with **4+ years** running production Kubernetes on **AWS, GCP** and hybrid on-premises infrastructure. Owns services end to end - **SLOs and error budgets**, on-call incident response, and the **CI/CD** that ships them. Applies the same reliability discipline to **ML and LLM platforms**, and contributes fixes upstream to the **CNCF** tools running in production. **Kubestronaut**.

TECHNICAL SKILLS

Reliability: SLIs (Service Level Indicators), SLOs (Service Level Objectives), Error Budgets, MTTR & MTTD Reduction, Incident Response, On-Call Rotation, Blameless Postmortems, Root Cause Analysis (RCA), Disaster Recovery (DR)

CI/CD & IaC: CI/CD (Continuous Integration / Continuous Delivery), GitHub Actions, GitOps, ArgoCD, Argo Rollouts, Argo Workflows, Jenkins, Blue-Green & Canary Deployments, Terraform, Infrastructure as Code (IaC), Ansible, Configuration Management

Cloud & Containers: AWS, GCP, Azure, Kubernetes (K8s), GKE, Docker, Podman, CRI-O, Linux (Rocky, CentOS)

Observability: Prometheus, Grafana, Loki, Tempo, Thanos, Alertmanager, PagerDuty, Custom Exporters, Distributed Tracing, Self-Healing Automation

ML Platform (MLOps / LLMOps): Model Serving & Inference, Drift Detection, GPU Node Scheduling, Continuous Training (CT), Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), Vector Databases, Agentic AI, LLM Observability, Tracing & Evaluation, Token & Cost Attribution, MLflow, LangGraph, LangSmith, vLLM, Ollama, FAISS, Temporal

Security & Data: HashiCorp Vault, Trivy, OWASP Dependency Check, Bastion & VPN Access Control, MySQL, MariaDB, PostgreSQL, ClickHouse, Redis, Trino, Apache Doris, Flink

Languages: Python, Go, Bash, SQL, YAML

EXPERIENCE

Senior DevOps Engineer | BrightEdge

Remote | Jan 2025 - Present

- Drove **deployment failures down 25%** and **deploy success up 60%** with blue-green **Argo Rollouts**, automated regression gates and one-command rollback.
- Modernised **300+ production machines** from **CentOS 8** to **Rocky Linux** across **GCP, AWS** and on-premise with an **Ansible**-driven workflow validating each staged batch.
- Architected an **agentic AI platform** on **LangGraph, Temporal, PostgreSQL, Redis** and **LangSmith**, removing **250+ manual hours per month**.
- Instrumented the agentic platform for **LLM observability** - **LangSmith** tracing on every agent span, with token and cost attribution per workflow and replay of failed tool calls for evaluation.
- Standardised provisioning on **Terraform** as the single path to production - **150+ MySQL VMs** and **10 GKE clusters** running **ClickHouse, Flink, Envoy** and **Prometheus/Loki/Grafana**.
- Cut **MTTR** on high-severity incidents by defining **SLIs, SLOs and error budgets** with burn-rate alerting that gates risky releases.
- Resolved a production egress outage where pods stayed **Running** and passed health checks yet intermittently failed outbound calls - traced to **Cloud NAT** port exhaustion at peak load and cleared in **under 45 minutes** by expanding the gateway pool **2 to 8 NAT IPs (64,512 ports each)**.
- Restored a stalled **48-node** cluster expansion during **pod IP exhaustion** - the new subnet's secondary range sat outside the **VPC firewall** rules, leaving **60+ pods Running but unable to reach the kube-apiserver**; admitting the range to the control-plane rules recovered scheduling in **~30 minutes**.
- Ran **15+ live production failovers** with zero data loss via **GTID** cutover and cross-zone static-IP reassignment, hardened after root-causing a **DNS** failure in the promotion path.
- Eliminated production blind spots with observability at **100% service coverage** on **Prometheus, Grafana, Tempo and Thanos**, plus custom **Go** exporters.

- Built **AIOps** self-healing across **175+ hosts** - **Prometheus** alerts correlated into automated remediation, plus swap-pressure detection that drains a node before restarting **MySQL**.
- Diagnosed production failures at depth on **GKE** - a **Trino** coordinator GC death spiral traced to an **Iceberg** write setting, and **ClickHouse Keeper** collection loss crash-looping 15 pods.
- Automated **ML training pipelines** on **Argo Workflows** with hyperparameter sweeps on cost-controlled **GPU nodes**, promoting **model code** through production **CI/CD** gates.
- Scaled **80+ orchestrated jobs across 1B+ records to 95% success** with retries and checkpointing, sizing queue-driven workloads to demand with **KEDA** autoscaling.
- Migrated **100+ cron jobs** off a single unmonitored VM onto **Argo CD** and **Argo Workflows**, replacing a silent scheduler with retries, run history and observability.
- Led migration of **25 MySQL clusters (~7,500 tables)** to **ClickHouse** via **Parquet** on **GCS**, deploying **Trino** on **GKE** and gating destructive truncation behind row-count verification.

DevOps Engineer | Sherlock AI

Mumbai, India | Jun 2022 - Dec 2024

- Improved service **availability 40%** by hardening **Kubernetes** workloads with **PDBs**, **HPA**, network policies and topology spread.
- Streamlined releases to **50% fewer deployment failures** and **40% faster deploys** with **GitOps CI/CD** on **ArgoCD** and automated manifest promotion.
- Reduced container images **80%** with multi-stage **Docker** builds, cutting pull time and attack surface.
- Established release gates with **Trivy** and **OWASP Dependency Check**, **Sealed Secrets** for credentials, and **Velero** backups with rehearsed restores.
- Operated **Apache Airflow**, **NiFi**, **PySpark** and **AWS EMR** clusters for the data science team.

PROJECTS

SLO-Gated Delivery Platform. Self-service **GitOps** platform that makes reliability a release gate: **ArgoCD** and **Argo Rollouts** step traffic through canary weights while **Prometheus** burn-rate queries score the **SLO error budget** at every step, pausing or reverting a rollout the moment the budget depletes. **Terraform** provisions the clusters, **Vault** issues the credentials, and **Thanos** retains the long-range signal the gates evaluate against.

Alert Correlation & Auto-Remediation Engine. **Alertmanager** webhook receiver that groups related firings into one incident, infers probable cause from recent rollouts and resource pressure, then runs a guarded remediation playbook that reverts itself on failure.

Agentic Workflow Runtime. Multi-step **LangGraph** agents executed as durable **Temporal** workflows - every step checkpointed to **PostgreSQL** so a failed tool call resumes instead of restarting - with **Redis** for shared state and **LangSmith** tracing each span for replay and evaluation.

Confluence RAG Assistant. Retrieval pipeline that chunks and embeds documentation into a **FAISS** vector store, then reranks candidates through **Ollama LLMs** before generation - covering chunking, embedding choice, semantic search and answer grounding.

Open Source (CNCF). Fixes merged upstream into projects including **Argo CD** (Flink health assessment), **Prometheus node_exporter**, **prometheus-operator**, **Flux**, **MetalLB** and **Dapr** - contributing back to the open-source tools I run in production by debugging upstream **Go** source.

EDUCATION

Bachelor of Engineering (B.E.), Computer Engineering - University of Mumbai, 2022 | CGPA 9.01/10

CERTIFICATIONS

- **Kubestronaut** - Cloud Native Computing Foundation (CNCF), awarded for holding all five Kubernetes certifications
- **CKS** - Certified Kubernetes Security Specialist
- **CKA** - Certified Kubernetes Administrator
- **CKAD** - Certified Kubernetes Application Developer
- **KCSA** - Kubernetes and Cloud Native Security Associate
- **KCNA** - Kubernetes and Cloud Native Associate
- **AWS Certified Solutions Architect - Associate**
- **ClickHouse Architecture**