I GEDE WIBAWA CAKRAMURTI (OCHA)

Platform Engineer · LLM inference and MLOps

Melbourne, Australia · Australian Permanent Resident · Open to remote

cakraocha@gmail.com · ochacakramurti.dev · github.com/ochacakramurti · linkedin.com/in/cakramurti

SUMMARY

Platform engineer specialising in LLM and ML inference. As Heidi Health's founding infrastructure engineer, scaled an AI medical scribe from its first AWS account to 5 data-residency regions on 3 clouds and led the security certifications hospitals and enterprises require. Now own Heidi's self-hosted speech-to-text platform, which replaced a managed vendor at lower cost, latency and error rate. Previously 4 years in Big 4 risk, audit and M&A.

SELECTED IMPACT

- Replaced a managed speech-to-text vendor with self-hosted Triton on EKS GPUs across all AWS production regions, cutting cost ~20%, p95 latency ~28% and errors ~7x.
- Launched 4 data-residency regions in 5 months (UK, Canada, US, EU) and led ISO 27001, SOC 2 Type II and Cyber Essentials certification, opening enterprise, hospital and NHS markets.
- Re-platformed the GPU fleet onto MIG-partitioned RTX PRO 6000 during a capacity shortage, raising throughput per node 3.4x and cutting cost per unit of throughput ~27%.

EXPERIENCE

Heidi Health (formerly Oscer), Melbourne Oct 2021 - Present

AI medical scribe used in 190+ countries, ~2.8M patient visits a week (Sep 2026).

Senior Cloud Engineer, Engineering Platform Aug 2025 - Present

- Designed and built the self-hosted speech-to-text platform (Triton, fine-tuned Parakeet, EKS GPU clusters, cross-region PrivateLink), taking it from prototype to production in one quarter.
- Led the vendor cutover through shadow traffic, a feature-flagged regional ramp and per-request fallback.
- Migrated the fleet from L40S to RTX PRO 6000 (g7e) with zero downtime, and rebuilt capacity management around reserved GPUs with Karpenter, NodeOverlay and automated reservation booking.
- Built alerting, runbooks and a progress-based liveness probe that recovers stalled Triton pods in ~14 minutes without paging; diagnosed and fixed shared-memory exhaustion and poison-input failures.
- Decommissioned the legacy ECS platform for 3 services with zero downtime, launched Azure production regions in Japan, the US and the UAE, and replaced broad AWS PowerUser access with 4 least-privilege PIM roles.

Acting Engineering Manager, DevOps Feb 2025 - Jul 2025

- Managed 3 engineers (junior, mid, senior); introduced Linear request intake, a ticketed help desk and a 4-round DevOps interview loop.
- Built the MongoDB Atlas platform and migrated terabyte-scale data off DocumentDB in 5 regions using DMS change data capture and a rollback path; cut search p95 latency ~80% and enabled vector search.

DevOps / Machine Learning Engineer, founding infrastructure engineer Jan 2022 - Jan 2025

- Owned infrastructure, CI/CD, databases, observability, security, compliance and IT as the sole infrastructure engineer, from the first AWS account through the Scribe launch (Oct 2023).
- Launched 4 data-residency regions in 5 months (UK, Canada, US, EU; 2024) and migrated existing users and data to their home regions.
- Led ISO 27001 (2023; 2022 transition 2024; 2025 audit), SOC 2 Type II (2024) and Cyber Essentials (2024) certification, largely single-handed; now supervise renewals.
- Established BCP and DR, asset tagging, incident response and customer security questionnaire processes.
- Designed the first AWS platform in Terraform, CI/CD for all backend services, a Go API Gateway authoriser and Datadog observability; ran company IT (laptops, MDM, network, DNS).

Machine Learning Engineer (intern to full time), Oscer Oct 2021 - Jan 2022

- Cut the patient/bot chat data, training and deployment cycle from ~50 to ~4 hours; built model CI/CD on MLflow.

Across all roles: interviewed backend system-design candidates and most platform team hires.

Earlier career

- **Cultural Infusion, Melbourne:** Junior Data Scientist · Nov 2020 - Apr 2021. Built a privacy risk assessment framework for a cultural diversity platform.
- **PwC, Jakarta:** Senior Associate, Risk Assurance Data & Analytics · Jan 2019 - Jan 2020. Automated a bank's country-risk reporting, cutting preparation time 80%; led a team of 2-4 with PwC Japan.

- **EY, Jakarta:** Consultant, Advisory, and Junior Auditor · Sep 2015 - Dec 2018. Implemented a risk management and internal audit framework for Bank Indonesia; trained 25 staff up to Deputy Director level.

TALKS

- **AWS APJ EKS Roundtable** (invite-only), Tokyo, Jul 2026: ECS to EKS migration and multi-region GPU inference under data localisation.
- **AWS re:Invent 2025**, Las Vegas: "How Heidi Health is leveraging GenAI to transform the global healthcare industry", solo lightning talk. Recording
- **AWS Startup Dev Day**, Melbourne, Aug 2025: MongoDB customer session on the DocumentDB to Atlas migration, with Kerry Zhang.

SKILLS

- **Inference and MLOps:** NVIDIA Triton, NeMo ASR, GPU Operator, MIG, DCGM, KEDA, Karpenter, Gateway API, MLflow
- **Cloud and platform:** AWS (EKS, Organizations, PrivateLink, IAM, DMS, Lambda), Azure (AKS, Entra ID), GCP (Vertex AI), Terraform, Helm, GitHub Actions, MongoDB Atlas
- **Reliability and security:** Datadog, OpenTelemetry, Prometheus, incident command, BCP and DR, ISO 27001, SOC 2, Cyber Essentials, Vanta
- **Languages:** Python, Go, Bash, SQL

EDUCATION

- **The University of Melbourne:** Master of Information Technology (Artificial Intelligence) · 2021. Engineering scholarship; academic tutor for the capstone software project, 2022.
- **Universitas Gadjah Mada:** Bachelor of Economics in Accounting · 2015. Cum laude.