Webomage: Engineering for the AI Era

Production-Grade Cloud & AI Infrastructure

Executive Summary

Webomage is a senior engineering consultancy dedicated to building resilient, scalable, and secure infrastructure for the next generation of SaaS and AI applications. We bridges the gap between traditional DevOps rigor and rapid AI product development, enabling companies to deploy "Day 2" operations on Day 1.

Our Philosophy: The "Production-Grade" Standard

We believe that modern infrastructure must be treated as a product, not a utility. Our specialized approach ensures:

- **Determinism**: Every environment, from dev to prod, is defined in code (IaC) and deployable with zero manual intervention.
- **Security by Design**: We do not "add" security later. Zero-Trust networking, Row-Level Security (RLS), and fine-grained IAM are our default starting points.
- **Observability**: You cannot fix what you cannot measure. We build deep instrumentation (Prometheus/Grafana) into the core of every platform.

Core Capabilities

1. Cloud-Native & Hybrid Infrastructure

We architect platforms that grow with your business, moving beyond rigid "lift and shift" migrations to true cloud-native implementations.

- **Multi-Cloud Engineering**: Seamless deployments across AWS (EKS) and GCP (GKE, Cloud Run), leveraging the best services from each provider.
- **Platform Engineering**: Building internal developer platforms (IDPs) using tools like **Porter.io** and **Serverless Framework** to accelerate team velocity.
- **Hybrid & Bare-Metal**: Cost-effective virtualization strategies using **Proxmox** and Terraform for workloads that demand dedicated hardware performance.
- **GitOps Pipelines**: Automated, audit-trail verified deployments using ArgoCD or Flux, ensuring the "Golden State" is always live.

2. AI & LLM Infrastructure

We help engineering teams integrate Large Language Models into their products without compromising on privacy or performance.

- **Enterprise RAG Pipelines**: Building scalable retrieval systems using **Haystack**, LangChain, and Supabase Vector for context-aware applications.
- **Cloud & Edge AI**: Leveraging **GCP Vertex AI** for managed pipelines (AutoML, STT/TTS) and **Supabase Edge Functions** for low-latency inference.
- **Inference Optimization**: Deploying and scaling open-weights models (Llama 3, Mistral) on your own hardware or efficient cloud instances using vLLM and MLX.

3. Serverless & Edge Computing

For applications requiring global scale and millisecond latency.

- **Edge-First Architecture**: Moving compute closer to the user using Cloudflare Workers and Supabase Edge Functions.
- **Global State**: Utilizing distributed databases and smart caching strategies to ensure data consistency across regions.

Case Study: "Project Webomage 7" (Internal Dogfooding)

To demonstrate our capabilities, we built our own platform using the exact stack we recommend to clients.

- **Challenge**: Migrate a legacy static site to a dynamic, AI-powered platform with zero maintenance overhead.
- **Solution**: Rebuilt on **Astro + Svelte** (Edge Frontend) and **Supabase** (Serverless Backend).
- **Result**:
  - **Performance**: Perfect "100" Lighthouse scores across the board.
  - **AI**: Integrated "Chat with Docs" RAG feature directly into the UI.
  - **Automation**: Unified Terraform pipeline orchestrating both edge functions and frontend builds.

Why Partner with Webomage?

We are not a staffing agency. We are a specialized engineering partner. When you work with us, you get:

- **Senior Expertise**: Direct access to engineers with 20+ years of experience.
- **No Vendor Lock-in**: We build on open standards (Kubernetes, Terraform, Docker).
- **Audit-Ready Deliverables**: Documentation and compliance artifacts provided by default.

Build the infrastructure your AI product deserves. *Contact us for a technical audit.*