Santiago Velázquez

GitHub | LinkedIn | santivelazqueziusim@gmail.com | +972-54-467-9320

Profile

DevOps and Platform Engineer with 3 years of hands-on experience designing, building, and operating production systems at scale. I focus on end-to-end ownership across infrastructure and backend platforms, from architecture and automation to deployment, monitoring, and incident response. Strong background in Linux, Kubernetes, AWS, networking, and Python-based backend development, with practical experience building ML- and LLM-backed systems that run reliably in production.

Experience

DevOps / Platform Engineer – Matzov Unit, Israel Defense Forces 2023 – 2025

- Designed, built, and operated high-availability Linux-based production systems supporting thousands of users.
- Designed and provisioned cloud and hybrid infrastructure using Terraform, including AWS networking, compute, storage, and IAM.
- Built, deployed, and operated Kubernetes (OpenShift) clusters using Docker and Helm, managing lifecycle, scaling, and rollouts.
- Designed and implemented Python/FastAPI backend stacks serving as the control plane for ML pipelines and LLM inference services.
- Deployed and operated ML workloads and LLM inference services in production, integrating them with backend APIs and CI/CD workflows.
- Built and maintained CI/CD pipelines using GitLab CI/CD for application, ML, and infrastructure deployments.
- Configured and operated networking components including ingress, load balancing, TLS, DNS, firewalls, and reverse proxies (NGINX, HAProxy).
- Implemented monitoring and observability using Prometheus and Grafana to track system and service behavior in production.
- Took full ownership of systems from architecture and environment design through production rollout and incident response.

Projects

- The Embedder – End-to-end Retrieval-Augmented Generation platform handling ingestion, chunking, indexing, and query serving via FastAPI, integrating Qdrant with local, Docker, Kubernetes (Helm), and AWS deployments.
- The Chunker – Python-based semantic chunking engine for source code and text, supporting AST-based chunking via Tree-sitter, token-aware segmentation, and overlap strategies optimized for embedding models.

Skills

Platforms & OS	Linux, AWS (EC2, ECS, VPC, IAM, S3), Kubernetes (OpenShift), Docker, Helm
Backend & ML	Python, FastAPI, REST APIs, LLM inference services, RAG pipelines
IaC & Automation	Terraform, Ansible, Bash
CI/CD & Observability	GitLab CI/CD, Prometheus, Grafana
Networking & Data	TCP/IP, DNS, HTTP/HTTPS, NGINX, HAProxy, PostgreSQL, Redis, Qdrant

Education

DevOps Engineer Program, Matzov Unit, Israel Defense Forces 2023 – 2025
Mechinat Melach Haaretz, Youth Leadership Program 2021 – 2022
Magshimim Cyber Program, National Cyber Education Initiative 2018 – 2021