Alberto Quintero

ML / AI Infrastructure Engineer

alberto@quintero.cc | quintero.cc | linkedin.com/in/ajquintero

Summary

Cloud engineer specializing in ML/AI infrastructure: GPU inference platforms on AWS, multi-region capacity and cost-efficient scaling. 17+ years in software engineering, with a computer vision background. I build the infrastructure as code, pipelines and observability that keep AI workloads reliable and affordable.

Skills

Cloud: AWS (compute, containers, serverless, networking, storage, data, ML, security) / GCP

ML/AI infrastructure: GPU and CPU inference on ECS, SageMaker, multi-region Spot capacity, throughput and cost-per-inference optimization, safety guardrails for generative AI features

IaC: Terraform / OpenTofu

CI/CD and containers: GitHub Actions / GitLab / Docker

Observability: Datadog / Prometheus / Grafana

Cost and security: FinOps (right-sizing, Spot, Savings Plans), least-privilege IAM, encryption

Languages: Python / C++ / Bash

Experience

Grand Shooting March 2026 – Now
Cloud Engineer *Brands / Retail / Marketplaces*

- Own the **AWS** infrastructure of the AI image-processing platform: GPU and CPU workloads on **ECS** across multiple regions, serverless components with **Lambda** and **SQS**
- Define infrastructure as code with **Terraform** / **OpenTofu**: multiple environments, remote state, reusable modules
- Build **CI/CD** pipelines for services and infrastructure: **GitHub Actions**, semantic versioning, automated deployments to **ECS**
- Work across **AWS** and **GCP**; designed the plan to migrate hosted environments (**IONOS**, **OVH**) to **AWS** and executed part of it
- Set up **observability** with **Datadog**: logs, metrics and monitors
- Reduce cloud cost: right-sizing, **Spot** capacity and scale-to-zero schedules for GPU services
- Apply **security** best practices: least-privilege IAM, encryption at rest and in transit

Diffusely (formerly Meero) October 2022 – February 2026
Cloud Engineer *Real Estate / Fashion / Cars*

- Designed and implemented the architecture of several client-facing systems
- Designed the architecture for serving ML models on **AWS SageMaker** (hosted and serverless endpoints)
- Developed reusable **Terraform** modules (**IaC**) for the company's infrastructure
- Defined **security** best practices, documentation and automated checks across modules
- Ran the on-premises data center with **Ansible** and **Docker**, with monitoring on **Prometheus** and **Grafana**
- Reviewed and proposed **MLOps** best practices company-wide

Diffusely (formerly Meero) July 2020 – October 2022
R&D Team Leader *Real Estate / Fashion / Cars*

- Led the integration team (~6 people): roadmaps, project follow-ups and career paths
- Coached the team on dockerization, code robustness, AWS integration and systems architecture
- Led **computer vision and image processing** projects, and bridged the R&D and IT teams
- Designed and ran the workflow that deploys all R&D projects to **production**

View Factor November 2018 – July 2020
Technical Founder *Traffic Control*

- Designed the **system architecture** for on-site vehicle and pedestrian detection on embedded systems, using deep learning models to control traffic lights at road intersections: **C++**, **OpenCV**, **TensorFlow**, **TensorRT**, **Python**, **Flask**
- Built a highway average-speed control system using automatic license plate recognition, with synchronization across remote nodes and a dashboard
- Built a body temperature screening application combining **thermal and RGB cameras** (**C++**, **Qt**), with face recognition and per-person temperature history

Fitle October 2014 – July 2018
Lead R&D Engineer *Fashion*

- Led **computer vision** R&D (1 engineer and 1 intern): planned and coordinated the short-, medium- and long-term roadmap
- Integrated a **deep learning** segmentation method (**Caffe**), which improved segmentation accuracy up to 95%+
- Designed and developed a tool for interactively deforming the pose of registered **3D** bodies
- Proposed and developed approaches combining superpixel segmentation, background subtraction and pose estimation
- Proposed and developed approaches for **3D reconstruction**, internal tools and mobile applications, with a focus on user experience and optimization
- Integrated a bundle adjustment library for camera calibration and 3D reconstruction into the initial prototype

Earlier experience

- **Software Engineer**, TMH International Group (Insurance), Sep 2011 – Jul 2012: redesigned the database and module organization of an electronic medical record system; queries went from 30+ seconds to a few milliseconds.
- **Research Assistant**, Universidad del Zulia, May 2010 – Sep 2011: taught introductory computer science; helped set up the university's knowledge transfer, innovation and entrepreneurship unit (roles and system architecture).
- **Software Engineer**, CIDTEC (Research Lab), Jun 2009 – Dec 2009: built a CMS and an editorial management system in PHP (OOP) for a scientific journal.
- **Software Engineer Intern**, CorpoElec (Electricity Supplier), Apr 2009 – Jun 2009: built a CRUD module over an Oracle database (PHP + JavaScript); ported an ASP.NET system to PHP + JavaScript.

Education

Master of Science from Université de Bourgogne 2014
Specialization in Computer Vision and Robotics

Bachelor of Science from Universidad del Zulia 2011
Specialization in Computer Science

Certifications

- AWS Solutions Architect - Associate