

● Open to new opportunities

Constantin Petruț

Kubernetes. Terrafo

DevOps engineer with 5 years in production infrastructure, pivoted hard into AI/ML 2 years ago. Now building ML pipelines, agentic AI systems, and computer vision services at DeviDevs, while running enterprise cloud infra at btProvider. Kubernetes, Terraform, Python. Production, not tutorials.

5

YEARS DEVOPS

2

YEARS AI/ML

3

CLOUD PLATFORMS

10

AI AGENTS BUILT

CORE COMPETENCIES

What I bring to the table

Deep expertise across the full MLOps lifecycle, from infrastructure provisioning to model serving in production.



MLOps & AI Infrastructure

Kubernetes

Argo Workflows

MLFlow

ArgoCD

GPU Scheduling

Model Serving

Terraform

Data Pipelines

Grafana

Prometheus



Cloud & Platform Engineering

AWS EKS

EC2

S3

Lambda

GCP GKE

BigQuery

Azure AKS

Docker

Helm

GitHub Actions

GitLab CI



AI & Development

Python

FastAPI

HuggingFace

OpenAI

Anthropic

YOLO

ResNet

Agentic AI

Node.js

Web3

WORK EXPERIENCE

Where I've made impact

From founding an AI company to operating enterprise infrastructure at scale.

ML INFRASTRUCTURE & MLOPS

- Designed and maintained cloud-based ML infrastructure on AWS and GCP for large-scale model training, experiment tracking, and data management
- Built automated ML pipelines for data preprocessing, model training, validation, and deployment using Argo Workflows on Kubernetes
- Implemented experiment tracking and model registry using MLFlow, enabling reproducible training runs and versioned model artifacts
- Operated Kubernetes clusters with GPU node scheduling for compute-intensive ML training jobs, managing resource quotas and autoscaling
- Architected model serving infrastructure with automated rollout, health checks, and rollback capabilities

AI & COMPUTER VISION

- Deployed and served computer vision models (YOLO, ResNet, custom classifiers) from HuggingFace for image classification, object detection, and content analysis
- Built inference pipelines with preprocessing, model serving, and post-processing stages for real-time and batch predictions
- Designed multi-agent AI systems with autonomous decision-making, tool orchestration, and iterative reasoning capabilities
- Built autonomous crypto trading agents with real-time market data processing, technical analysis, and risk-managed execution
- Integrated multiple ML providers (OpenAI, Anthropic, Google, HuggingFace) with load balancing and fallback routing

INFRASTRUCTURE AS CODE & CI/CD

- Managed all cloud infrastructure using Terraform with modular, reusable configurations across AWS and GCP
- Built CI/CD pipelines (GitHub Actions, ArgoCD) for automated model training triggers, validation gates, and production deployment
- Containerized all ML workloads with Docker, deployed on Kubernetes with Helm charts
- Monitoring and alerting for model performance, data drift detection, and infrastructure health using Grafana and Prometheus

- Production ML systems processing thousands of inference requests daily with 99.5%+ uptime
- Reduced model deployment time from days to hours through automated ML pipeline orchestration
- Delivered scalable API-first platforms and data processing engines for enterprise clients

Python

FastAPI

Docker

Kubernetes

HeLm

Terraform

AWS

GCP

ArgoCD

Argo Workflows

MLFlow

HuggingFace

Grafana

Prometheus

Web3.py

- Production cloud infrastructure management on AWS and GCP at scale (high availability, security, performance)
- Container orchestration with Docker and Kubernetes: workload scheduling, scaling, resource optimization
- Infrastructure as Code with Terraform and CloudFormation for auditable, consistent provisioning
- CI/CD pipeline design and maintenance in GitLab for multi-environment automated deployments
- Scalable data pipelines and ETL workflows in Talend, automating data processing across enterprise systems

- Production API development in MuleSoft for high-throughput system interoperability
- Developing AI-powered automation agents for workflow optimization and intelligent infrastructure monitoring
- Real-time monitoring with Grafana, security audits, disaster recovery automation
- Mentoring junior engineers on DevOps practices and cloud infrastructure

AWS

GCP

Kubernetes

Docker

Terraform

GitLab CI

Talend

MuleSoft

Grafana

Python

vidaXL

2021 – 2022

📍 Bucharest, RO

Senior System Engineer

- Production system monitoring and incident response for global e-commerce platform
- Virtualization infrastructure management (VMware) for large-scale operations
- Data transformation pipelines between Salesforce and SAP via ESB
- Major incident escalation point, cross-team coordination and post-mortem analysis

Ericsson

2020 – 2021

📍 Bucharest, RO

IoT Platform Engineer

- IoT platform monitoring, troubleshooting, and edge computing operations
- Incident management for customer-facing IoT deployments

NOTABLE PROJECTS

Built to run in production

Systems I've designed and deployed, handling real workloads every day.

MLOps

End-to-End ML Pipeline Platform

Fully automated ML pipeline: data ingestion, preprocessing, training, validation, model registry, and serving. MLFlow for experiment tracking, Argo Workflows for orchestration on Kubernetes. Automated retraining triggers on data drift detection.

Argo Workflows

MLFlow

Kubernetes

Python

Computer Vision

CV Inference Service

Production deployment of HuggingFace vision models for real-time image classification and object detection. Kubernetes-based serving with GPU scheduling, autoscaling, and health monitoring.

HuggingFace

YOLO

Kubernetes

FastAPI

Web3 / AI

Autonomous Crypto Trading Agent

AI-powered trading system with real-time market data ingestion from multiple exchanges, technical analysis pipeline, ML-based

Agentic AI

Multi-Agent AI Orchestration

Autonomous AI agent platform with 10+ specialized agents collaborating on complex workflows. Self-learning feedback loops,

signal generation, and risk-managed order execution with automated portfolio rebalancing.

Python

Web3.py

WebSocket

Exchange APIs

dynamic tool orchestration, and multi-provider LLM routing. Thousands of autonomous decisions daily.

Claude

OpenAI

Python

Custom Framework

EDUCATION

Foundation



Universitatea Tehnica de Constructii Bucuresti

Engineering Degree

Let's build something great

Looking for an MLOps engineer, AI developer, or DevOps specialist? Let's talk.

✉ contact@devidevs.com

☎ +40 740 462 388

🌐 [LinkedIn](#)

🌐 devidevs.com