ALEXANDRU CIOC

Software Engineer · AI agents, Java/Spring Boot, Python · Former CTO, Leyoda

Based in Spain · EU citizen, no visa or sponsorship needed · Open to remote, hybrid or relocation

+40 742 211 051 · alex@ciocandco.com · github.com/WhitehatD · linkedin.com/in/alexandru-cioc-a40a3b247

PROFILE

Software engineer and former CTO of Leyoda, an AI company, where I built and ran the platform end to end. I work across the whole stack, from Java and Python backends to Kubernetes, and build the AI agents and RAG on top. Coding agents write most of my code, and my job is to steer them and check that what comes back holds up.

EXPERIENCE

Leyoda · CTO & Technical Co-founder

Oct 2025 – Oct 2026

AI that runs on the customer's own hardware · leyoda.eu

- Set the company's technical direction and the standards for how we built and shipped software
- Built the platform end to end in **Java 21 / Spring Boot** and **PostgreSQL**, with every frontend in **Next.js**, designed as modular services on one identity layer and data model
- Built our own **login and identity server (OIDC)**: one sign-in across all products, service-to-service access and revocable sessions
- Built a CRM with an **AI assistant** that works through 25 logged, permission-checked tools, and opened the same tools to Claude and Codex over **MCP** under the same rules
- Planned an agent for our first client, an accountancy firm, that works inside the software they already use, with the models running on hardware in their office (**llama.cpp**)
- Owned the infrastructure and zero-downtime deploys for three production services, plus SEO and AI-search visibility for the company's sites
- Wrote the pricing, unit economics and hiring plan behind the seed round

MetricHost · Platform Engineer (independent contract)

Dec 2025 – Jun 2026

Game-server hosting platform built for Atlas Studios, with an engineering partner · github.com/WhitehatD/MetricHost

- Designed and built it from a one-page concept to a production-ready system: 10+ **Spring Boot** services and **Kafka** on **Kubernetes** clusters in the EU and US, set up with **Terraform** and **Ansible**, with **Stripe** billing and monitoring in Prometheus, Grafana and Loki
- Cut the cost of idle game servers to about a third by putting them to sleep in under a second, and CI time from 25 minutes to 8

Outlier (Scale AI) · AI training, coding tasks

2025 – 2026

Part-time contract · wrote and reviewed programming tasks used to train AI code models

Freelance · Java Backend Developer

2020 – 2025

Client work for Vertmix (agency), Piki Studios (studio) and EndPixel (lead developer), on multiplayer game backends with paying players

- Shipped 50+ Java services and plugins to live multiplayer servers: game modes, economies with shops, banks and player-owned businesses, progression systems and world-editing tools
- Joined several servers into one network: shared state over **Redis** pub/sub, players moved between servers through a proxy, and **MongoDB** writes on async thread pools so the database never blocked the game loop
- Built ranked matchmaking with ELO ratings, a search window that widens while a player waits, ping limits and thread-safe queues; each match started an arena on another server and moved the players over
- As lead developer at EndPixel, introduced Redis messaging between servers so the network could scale out

SELECTED PROJECTS

- **Hesperus** (bachelor thesis, presented at **BNAIC/BeNeLearn 2026**): an AI agent that plans when a low-power camera takes photos and checks each one with a vision model, built from **STM32** firmware to a **FastAPI** backend and Next.js dashboard. The agent runs on **LangGraph** and **LangChain**, with **RAG** over earlier photos. Compared 6 vision models to pick one. `github.com/WhitehatD/hesperus`
- **On-call agent for AWS** (open source): when a CloudWatch alarm fires, a **LangGraph** agent investigates through read-only AWS APIs and proposes one fix with its undo. The fix runs only after a human approves it and is reverted if the alarm does not clear. The limits are enforced by **IAM**, not the prompt. Built on **Step Functions**, **Lambda**, EventBridge and DynamoDB with **Terraform**. Found the right cause and fix in 10 of 10 failure scenarios, including a prompt injection. `github.com/WhitehatD/oncall-agent-aws`
- **KYB review agent** (open source): an agent that checks business customers against a company registry, sanctions lists and their documents, and earns more autonomy in stages (shadow, suggest, execute) only by passing evals on cases it never saw. Each run gets a signed, short-lived token, and the tools sit behind an **MCP** server that checks it on every call. Hard rules run in code after the model, every case is one **OpenTelemetry** trace across agent and tool server, and an Operator dashboard shows the queue and audit log. `github.com/WhitehatD/kyb-review-agent`
- **crag** (open source): keeps one set of rules for every coding agent in a project and checks them in CI and against the real files and commands. `github.com/WhitehatD/crag`

SKILLS

Languages: Java, Python, TypeScript/JavaScript, Go, C, SQL, Bash

Backend: Spring Boot, FastAPI, Node.js

Frontend: React, Next.js, Tailwind

AI: AI agents, LangGraph, LangChain, MCP, RAG (BM25 + embeddings), agent evals, llama.cpp, vLLM, Claude Code, Cursor, Codex

Cloud: AWS (Lambda, Step Functions, EventBridge, DynamoDB, SQS, API Gateway, IAM, CloudWatch), Hetzner

Infrastructure: Kubernetes, Docker, Terraform, Ansible, Helm, Linux, nginx, Cloudflare

CI/CD: GitHub Actions, self-hosted Forgejo, blue-green deploys, gated production releases

SEO / AI search: technical SEO, structured data (JSON-LD), crawler policy, sitemaps and IndexNow

Data and monitoring: PostgreSQL, Redis, Kafka, MongoDB, Prometheus, Grafana, Loki, OpenTelemetry

Testing: JUnit, pytest, Testcontainers

EDUCATION

BSc Computer Science, Maastricht University · thesis done and presented at BNAIC 2026, two exams left 2023 – 2026

Languages: Romanian (native), English (fluent, C2)