

RAMI OSMAN

Solutions Architect & Tech Lead — Cloud · Distributed Systems · Applied AI / GenAI

www.osmanrami.fr · me@osmanrami.fr · 06 02 73 49 22 · Paris

PROFILE

Solutions Architect & Tech Lead (8+ years) specializing in Cloud, distributed architectures and applied AI / GenAI, with experience on large-scale, mission-critical platforms for major enterprises (ENSO, SNCF, APHP, BNP, Volkswagen, Inserm). Currently designing an enterprise LLM platform end to end — secure model access, RAG, agentic actions with human-in-the-loop validation, full auditability and LLM evaluation gates — while leading technical teams and complex delivery.

EXPERIENCE

AI Solutions Architect — ENSO · 2026 – Present · Paris

Enterprise AI Platform (GenAI Programme)

Lead architect of a shared, multi-tenant AI platform powering four product capabilities (payroll declaration correction assistant, timesheet OCR, in-product assistants, autonomous document-chasing agent). Authored the Architecture Design Document and drive its enforcement across teams (architecture reviews, ADRs, compliance gates).

- Designed a central AI gateway as the single access path to Amazon Bedrock — EU-only inference profiles enforced organisation-wide (SCP), Guardrails PII filtering, model routing isolating products from model lifecycle.
- Built the RAG architecture on Aurora PostgreSQL + pgvector (HNSW, hybrid search + reranking) behind an owned retrieval API with a strict citation policy — every answer cites its sources or declines; tenant isolation enforced in-database via Row-Level Security.
- Specified agentic actions through a single MCP action server: versioned tool manifests, human validation gates on every write, idempotency and precondition checks, append-only decision log making every AI decision replayable (WORM export to S3 Object Lock).
- Zero-privilege-escalation identity design: the AI inherits the user's permissions via OAuth2 Token Exchange (Keycloak OIDC) and IRSA workload identities — never more.
- Quality & observability loop: DeepEval golden-set gates blocking regressions in CI; Langfuse self-hosted in-VPC on EKS for traces, prompts and online feedback; GitOps delivery with feature-flagged dark releases.
- Hybrid TypeScript / C# architecture (Mastra agent runtime, official MCP SDKs) with the language seam on the MCP protocol boundary; contract tests and a shared schema repository guarding the seam.
- GDPR & AI Act by design: EU data residency, data minimisation, coordinated right-to-erasure across vectors, traces and audit, AI transparency to end users.

Solutions Architect / Tech Lead — Inventiv IT · 2025 – 2026 · Paris

Inserm — Software Factory

Designed and implemented a company-wide software factory for 12 teams. Standardized CI/CD pipelines, improved code quality and reduced delivery times (GitLab Enterprise, Nexus, SonarQube, Kubernetes, Ansible, HDS-compliant data).

- Automated CI/CD pipelines for 40+ projects with HDS compliance.
- Kubernetes & Ansible industrialization reduced manual interventions by 50%.

Volkswagen Group — GENWIN Web Platform & SA Platform

Architecture and deployment of a highly available web platform on multi-AZ AWS infrastructure (IaC). Clean Architecture application design; guided teams on TDD, quality and CI/CD; functional scoping, architecture documentation (DAT) and Agile sprint follow-up with business teams.

SKILLS

AI / LLM Platforms

Amazon Bedrock (EU inference profiles, Guardrails) • RAG • pgvector / HNSW • Hybrid search & reranking • MCP (Model Context Protocol) • Agentic AI • Mastra / LangGraph • Human-in-the-loop validation • LLM evals (DeepEval, golden sets) • LLM observability (Langfuse) • Local LLM • Fine-tuning • ComfyUI • Vespa AI • n8n • TensorFlow • Databricks

Cloud / OPS

AWS • Multi-AZ architecture • EKS • SCP / IAM / IRSA • Step Functions / EventBridge • SageMaker • GCP (Cloud Run, PubSub) • Cloud Armor / VPC / Networking • Firebase • Terraform (IaC) • Ansible • Kubernetes • Docker • SOPS • GitLab CI/CD (Enterprise) • GitHub Actions • GitOps (Argo CD / Flux) • Feature flags (Unleash) • TeamCity • Nexus • SonarQube • Observability (Grafana) • HDS-compliant environments

Application Architecture

Hexagonal • Microservices • Clean Architecture • Multi-tenant (RLS) • Event-driven

Security & Compliance

Keycloak / OIDC • OAuth2 Token Exchange • GDPR & AI Act by design • Audit / WORM (S3 Object Lock) • PII filtering

Languages / Frameworks

Java • Spring Boot • TypeScript • Node.js • React / Next.js • C# / .NET • Dart • Flutter

Messaging / Queue

Kafka • AWS SNS / SQS • GCP PubSub

Databases

PostgreSQL (Aurora, pgvector, RLS) • Redis • MongoDB / Firestore

EDUCATION

EISTI (CY Tech) — Engineering Degree, specialization in Artificial Intelligence

Lycée St-Exupéry (CPGE) — Preparatory Classes MPSI / MP

Harbour.Space, Barcelona — Cyber Security programme / Linux systems

CERTIFICATIONS

AWS Partner: Generative AI Essentials
Codility Silver Award
SkillValue Java 11

LANGUAGES

French (native) • English & Arabic (fluent)

- Designed an AWS Multi-AZ HA architecture ensuring 99.9% availability.

SNCF — Data Market Place / MyCot Mobile Application

Designed a cloud-native data platform enabling large-scale data exposure and consumption; microservices on managed services with integrated observability and monitoring.

- Deployed Spring Boot microservices on GCP used by 90,000 users.

RATP — Antelope Platform Migration (Struts / Java 8 → Spring / Java 21)

Technical leadership of a legacy platform migration toward a modern Spring Boot microservices architecture. Experimented with and integrated AI agents to accelerate technical tasks and improve team productivity.

System Analyst — IBM France • 2023 – 2024 • Paris

Mars Inc. — Project 1Nav

Industrialized DevOps processes in an international environment; automated deployments and data workflows.

- Automated global DevOps processes reducing manual tasks by 45%.

APHP — E-Reply System / AI Models for Medical Follow-up

Designed an AI-augmented medical monitoring platform for automated analysis of patient communications. Integrated LLMs, RAG and vector search to improve response relevance and assist medical teams, in a highly demanding security and data-confidentiality environment.

- Implemented Vespa AI vector search reducing email response time by 70%.

BNP — Financial Instruments Account (CIF)

Development and evolution of a critical financial system based on hexagonal architecture and microservices.

Full Stack Developer — Diverta Inc. • 2022 • Tokyo

Kuroko: BaaS platform (Backend as a Service) / BFF (Backend for Frontends).

Full Stack Developer — Capgemini France • 2018 – 2022 • Paris

Stellantis — GCB Project (Bill of Materials Transformation FCA/PSA)

Transformation and harmonization of Bills of Materials between FCA and PSA systems in a large-scale industrial merger. Designed robust backend services ensuring data reliability, consistency and traceability within a complex multi-team information system.

GARI — Collaborative Platform

Freelance & Personal Projects • 2020 – 2025 • Paris

Leventis — Sales Prospecting Platform

Full-stack platform (Next.js, NestJS, Supabase) used across multiple business applications; shared APIs enabling multi-application integration; CI/CD fully automated via GitHub Actions.

Clairvoyance — Web & Mobile Application

Flutter + Firebase + GCP platform connecting experts and users, supporting 30,000 active users; real-time system, analytics-driven engagement, scalable and highly available cloud infrastructure.

Magik AI — Web & Mobile Application

Interactive media generation platform on open-source AI models, deployed on a RunPod GPU cluster for distributed inference; GenAI pipelines (RAG, fine-tuning, ComfyUI), automatic upscaling, AI agents and n8n workflows automating media production.