

Compétences

Cloud / OPS

AWS • Multi-AZ Architecture • SageMaker • GCP Cloud Run • Cloud Armor VPC • Networking • Firebase • Cloud Analytics • Terraform (IaC) • Ansible • Kubernetes • Docker • SOPS • GitLab CI/CD (Enterprise) • GitHub Actions • TeamCity • Nexus • SonarQube • DevOps Process Automation • Global DevOps Delivery • HDS-compliant • Data Environments Monitoring • Observability (Grafana)

Architectures Applicatives

Héxagonale • Micro services

Languages / Framework

Java • Spring Boot • TypeScript • React / Next JS • Dart • Flutter

Messaging / QUEUE

Kafka • AWS SNS / SQS • GCP PubSub

AI / LLM

Local LLM • RAG • Agentic AI • n8n • Tensorflow • Fine tuning • ComfyUI • Vespa AI • Databricks

Database

Relational • PostgreSQL In Memory • Redis Documents • MongoDB / Firestore

Formation

EISTI (CY TECH)

Diplôme d'ingénieur, spécialisé en intelligence artificielle

Lycée St Exupéry - CPGE

Classes préparatoires MPSI-MP

Harbour Space - Barcelone

Programme système linux et Cyber Sécurité

Certificats

AWS Partner: Generative AI Essentials

Codility Silver Award

SkillValue Java 11

RAMI OSMAN

PROFIL

Architecte Technique (7+ ans) spécialisé en Cloud, architectures distribuées et IA appliquée, avec expérience sur plateformes critiques grands comptes (SNCF, APHP, BNP, Volkswagen). Expertise en conception d'architectures scalables, sécurisées et en pilotage technique d'équipes et de projets complexes.

EXPÉRIENCES

Tech lead / Architect Solutions chez Inventiv IT — 2025 - Présent - Paris

• Inserm : Usine logicielle

Mise en place de l'usine logicielle, 12 équipes.

Standardisation CI/CD, Amélioration de la qualité du code et des temps de livraison Gitlab Entreprise, Nexus, SonarQube, CI / CD, Données HDS, Kubernetes, Ansible.

Résultat : Automatisation des pipelines CI/CD pour 40+ projets avec conformité HDS Industrialisation Kubernetes & Ansible diminuant les interventions manuelles de 50%

• Volkswagen Group : Plateforme Web GENWIN et Plateforme SA

Architecture et déploiement d'une plateforme web hautement disponible sur infrastructure cloud AWS multi-AZ, avec une approche Infrastructure as Code.

Mise en place d'architectures applicatives scalables et maintenables (Clean Architecture), accompagnement des équipes de développement sur les bonnes pratiques (TDD, qualité, CI/CD). Participation active au cadrage fonctionnel rédaction de documents d'architecture (DAT) et suivi des sprints en méthodologie agile, en lien direct avec les équipes métier.

Conception d'une architecture AWS Multi-AZ HA garantissant 99.9% de disponibilité.

• SNCF : Plateforme Data Market Place / Application mobile MyCot

Conception d'une plateforme data cloud-native permettant l'exposition et la consommation de données à grande échelle. Mise en œuvre d'architectures orientées microservices et services managés, avec intégration de l'observabilité et de la supervision.

Déploiement de microservices Spring Boot sur GCP utilisés par 90 000 utilisateurs.

• RATP : Migration Plateforme Antiloppe (Struts Java SE 8 vers Spring Java 21)

Pilotage technique de la migration d'une plateforme legacy vers une architecture moderne basée sur des microservices Spring Boot.

Expérimentation et intégration d'outils IA (agents) pour accélérer certaines tâches techniques et améliorer la productivité des équipes.

Analyste Système chez IBM France — 2023 - 2024 - Paris

• MARS INC : Projet 1Nav

Industrialisation des processus DevOps dans un contexte international, avec automatisation des déploiements et des flux de données.

Automatisation des processus DevOps globaux réduisant les tâches manuelles de 45%

• APHP : E-Reply System / Modèle IA de suivi médical par mail / plateforme

Conception d'une plateforme de suivi médical augmentée par l'IA, permettant l'analyse et le traitement automatisé de communications patients. Intégration de modèles IA (LLM, RAG) et de moteurs de recherche vectorielle pour améliorer la pertinence des réponses et assister les équipes médicales, dans un contexte exigeant en matière de sécurité et de confidentialité.

Implémentation Vespa AI Vector Search réduisant le temps de réponse aux mail

Langues

Français (Natif)

Anglais (Courant)

Arabe (Courant)

de 70%

- **BNP** : Compte d'instruments financiers (CIF)
Développement et évolution d'un système financier critique basé sur une architecture hexagonale et microservices.

*Développeur full stack chez **Diverta Inc.** — 2022 - 2022 - Tokyo*

- **Kuroko** : Plateforme BaaS (Backend as a Service) / BFF (Backend for frontends)
PHP 5, React, Google Cloud Platform, Github, MySQL.

*Développeur full stack chez **Capgemini France** — 2018 - 2022 - Paris*

- **STELLANTIS** : Projet GCB - Transformation bill of materials entre FSA et PSA
Transformation / harmonisation des Bill of Materials entre les systèmes FSA et PSA, dans un contexte de fusion industrielle à grande échelle. Conception et développement de services backend robustes assurant la fiabilité, la cohérence et la traçabilité des données, avec un fort accent sur la qualité logicielle, la performance et l'intégration dans un SI complexe multi-équipes.
- **GARI** : Plateforme collaborative

Projets freelance / personnels — 2020 - 2025 - Paris

Réalisation de plateformes web et mobiles de bout en bout, de la conception technique jusqu'à la mise en production. Interventions sur des projets orientés startup et innovation, autonomie sur les choix d'architecture, la stack technique et l'optimisation des performances. Développement de projets intégrant IA générative, agents intelligents et automatisation.

- **Leventis Platform** : Plateforme de prospection commerciale
Conception et développement d'une plateforme de prospection commerciale utilisée sur plusieurs applications métiers.
Développement Full Stack (Next.js, NestJS, Supabase) améliorant la productivité des équipes commerciales.
Mise en place d'APIs mutualisées permettant l'intégration multi-applications.
Industrialisation CI/CD via GitHub Actions automatisant 100% des déploiements.
- **Clairvoyance** : Application web et mobile
Développement d'une application mobile et web de mise en relation entre experts de la voyance et utilisateurs.
Architecture Flutter + Firebase + GCP supportant 30 000 utilisateurs actifs
Mise en place d'un système temps réel améliorant la fluidité des échanges utilisateurs
Intégration Google Analytics optimisant l'engagement et la rétention
Déploiement cloud scalable assurant une haute disponibilité de la plateforme.
- **Magik AI** : Application Web et Mobile
Conception d'une plateforme web et mobile de génération de médias interactifs basée sur des modèles IA open source.
Déploiement sur cluster GPU RunPod pour l'inférence distribuée
Implémentation de pipelines GenAI (RAG, fine-tuning, ComfyUI) réduisant le temps de génération média
Mise en place d'un système d'upscale automatique améliorant la qualité visuelle des contenus générés.
Orchestration d'agents IA et workflows n8n automatisant la production média