Marin Bouthemy

+65 8102 5931 | marin.bouthemy@gmail.com | Singapore
Status: Permanent Resident in Singapore | 1 Month Notice Period
LinkedIn: https://www.linkedin.com/in/marinbouthemy/
Github Portfolio: https://mbouthemy.github.io/blog/

SUMMARY

Experienced AI/software engineer with 6+ years of professional experience, including hands-on work building RAG systems, LLM-based conversational applications, and scalable ML infrastructure. Strong back-end foundation (Java, Python, API, PostgreSQL, Spark) combined with DevOps expertise (CI/CD, Kubernetes, Helm, Terraform) for deploying and scaling AI workloads in production. Passionate about applying modern AI techniques to solve real business problems, with a track record of reducing infra costs and latency at scale. Currently residing in Singapore with permanent resident status (no visa required).

WORK EXPERIENCE

Terrascope (3 years, 10 months)
Singapore

Sr. AI Engineer **Jan 2025 – Present**
- Built and deployed an in-house RAG (Retrieval-Augmented Generation) system using LLaMA, powering a MongoDB-backed conversational chatbot and reducing dependency on third-party LLM APIs
- Designed a microservices caching layer that cut RAG/LLM API response time by 93% (7s → 0.5s) at scale
- Orchestrated ML/LLM workloads on Kubernetes with Karpenter for dynamic node scaling, and automated provisioning with Terraform (IaC)
- Led migration to a multi-tenant architecture, reducing cloud costs 90% ($200k → $30k/month), including infra supporting AI workloads
- Supported ISO 27001 audits, including controls around data handling for AI/ML systems

Senior Data Scientist **July 2022 – Jan 2025**
- Built the backend (FastAPI, Actix, MongoDB, Kafka) of different microservices for users recommendations and forecasting in a modular and scalable way.
- Deployed the whole solution with Kubernetes / ArgoCD and Github CI/CD. Usage of Helm Chart and Terraform to automate the deployment process.
- Deployed the solution on **more than 30 clients instance** of AWS / Azure and integrated with Cloud components (Databases, event hubs messaging, Kafka...).
- Created the frontend website with React and Next to display results and demonstrate the usage of the different microservices before they get productionized.

Tools: Kubernetes, ArgoCD, Github CICD, Terraform, Kafka, FastAPI, Actix, MongoDB

SCOR **Feb 2020 – July 2022**

Data Scientist

Singapore

• Executed the creation of the backend (FastAPI, PostgreSQL, Kafka) and frontend of the application for an Optical Character Recognition (OCR) and Natural Language Extraction (NLP) service as an micro-services architecture with Kafka messaging**.**

• Deployed the whole solution with Kubernetes and Gitlab CI/CD. Usage of Helm Chart.

• Deployed the solution on client instance of AWS / Azure and integrated with Cloud components (Databases, event hubs messaging, Kafka...).

• Utilisation of Spark to store and extract logs information and statistics about the document and words extracted.

• Created the frontend website with React and Next to display results.

• Conducted workshops on coding best practices and on fairness and ethics in AI

• Coached and supervised interns and students for research projects

• Passed and validated multiple SOA exams (5/10 in less than 2 years)

The Boston Consulting Group **Jun 2019 – Jan 2020**

Visiting Data Scientist *Amsterdam, The Netherlands*

• Consultant for an international airline company, developing mathematical modeling and software development for optimal decision making

• Development and deployment of a Ground Services task allocation (**Mixed Integer Programming**) written with **Python and Gurobi**.

• Setting-up of a CI/CD pipeline with Docker and Kubernetes.

• Participated in discussion with stakeholders, displaying strong oral and written skills

Tools: Spark, Gurobi, Python, Docker, Kubernetes.

OpenClassrooms **Sep 2020 – Present**

Mentor in Python and Data Science *Remote*

• Weekly mentoring of professionals seeking to upskill themselves

• Mentored more than 15 professionals in total, ensuring best coding practices are respected, answered technical questions and provide assistance throughout their curriculum.

Barclays **Jun 2018 – Sep 2018**

Developer — Summer Intern *Glasgow, United Kingdom*

• Worked cooperatively in a team to design and develop an Acceptance Test Framework for Transaction Reporting applications

• Planned and delivered a scalable web application for application support team as a proof of concept

Tools: Java, Cucumber, WireMock, Maven, React

Seoul National University **Jun 2017 – Sep 2017**

Visiting Research Intern *Seoul, South Korea*

• Coded data science models for Quantitative Finance (Long Short Term Memory Networks)

• Implementation of Deep Reinforcement Learning and Policy Gradient for trading algorithms

EDUCATION

ENSAE ParisTech (Paris) **Sep 2015 – Jun 2019**
Master of Science in Data Science *Paris, France*
• Relevant Coursework: Dynamic Optimization, Numerical Methods and Applications, Algorithms Theory, Java and C
• Academic Project: Monte-Carlo Simulation with Gibbs Sampling and Metropolis-Hastings, Memory Efficient Kernel Approximation, Anomaly Detection in Networks

Hong Kong University of Science and Technology **Sep 2016 – Jan 2017**
Master of Science in Big Data – Visiting Student *Hong Kong*
• Relevant Coursework: Parallel Programming, Statistical Prediction, Applied Data Analytics

Mines Saint-Étienne **Sep 2015 – Jun 2019**
Master of Science in Computer Science and Mathematics *France*
• Relevant Coursework: Java, C, Shell, OCaml, JavaScript, Parallel Programming (OpenCL, MPI, CUDA), Algorithms & Data Structure

PROJECTS AND AWARDS

Web2To3 – First Prize – Singapore Blockchain Innovation Program 2022
• Created a React package called Play2Earn which allows web2 developers to easily transform their game into a play2earn game on Solana blockchain. Created the backend API with Rust, Actix Web and used Nextjs, React to code the interface for developers to interact with.

Participation in the Pytorch Summer 2020
• Creation of an mobile application incentivize kid to clean garbage using React Native. Implemented an algorithm for detection of garbages in pictures with EfficientNet and created the backend with Django and Pytorch.

Candy Machine Deployment on Solana Blockchain
• Deployment of NFT machine for a patreon NFT art project. Used Metaplex and Nextjs for website.

Chat Meeting Application
• Created with React Native and Firebase. Developed a chat application only based on users voices. Deployed the application on Google and Apple Store.

Generation Machine Learning - 3rd Prize (2018)
• Creation of a machine learning algorithm for fraud detection based on user tastes. Organized by Société Générale and HSBC in France.

Datathon 2nd Prize (2017)
• Ernst & Young 24 hours Deep Learning Competition in Paris.

VOLUNTEERING AND TEACHING

Saturday Kids Singapore

• Teach programming languages (Scratch & Python) for underprivileged kids in the local Singapore community.
• Explain programming concepts and coach the creation of video games using Scratch and Strawbees.
• Mentored more than 30 students from 8 to 14 years old.

Go Camp Organization
• Volunteering as a French school teacher in Ukraine for a month

Alumni Guest Lecturer
• Gave series of Lecture about Ethics & AI for university students

PROGRAMMING, SKILLS AND INTERESTS

• **Programming Skills:** Googling / StackOverFlow (advanced), Python (advanced), Java (advanced), Typescript (advanced), React / Next (advanced), Docker (advanced), Kubernetes / Helm (advanced), R (advanced), SQL, Ocaml (good), Angular (good), React Native (good), Rust / Solana (good)
• **Languages:** French (*native*), English (*fluent*), Mandarin Chinese (*HSK 4*)
• **Interests:** Running & Trekking, Hiking (Climbed Kilimanjaro in 2020), Classical Literature, Video Editing