

Bijon Setyawan Raya

bijonsetyawan@gmail.com

github@bsraya · <https://bijonsetyawan.com>

EXPERIENCE

Senior Machine Learning Engineer

October 2025 – Now

IKG Team Ltd.

Taipei, Taiwan

- Designed and led the end-to-end development of a real-time Content Moderation platform using Go, Redis, and MongoDB for similarity searches to decrease response time <1s and token consumption up to 90%.
- Provisioned a scalable Data Lakehouse on AWS using Terraform and orchestrated data transformation pipelines with Apache Airflow, establishing a tiered architecture (bronze, silver, gold) for real-time business analytics.

Machine Learning Engineer

November 2024 – October 2025

Cathay United Bank

Taipei, Taiwan

- Engineered a centralized LLM Gateway on AWS to consolidate enterprise access, implementing Guardrails that eliminates 99% of unsafe interactions while reducing latency by 90%.
- Conducted deep-dive profiling of bare-metal vs. virtualized GPU environments, identifying key bottlenecks in vLLM deployment to optimize resource utilization and stability.

Intermediate Fullstack Developer

October 2022 – March 2024

Faria Education Group

Taipei, Taiwan

TECHNICAL SKILLS

Languages: C/C++, Golang, Python*, SQL, Zig.

Deep Learning & Machine Learning: CUDA, cuML, HIP/ROCm, MLflow, PyTorch*, Scikit-learn.

Data Engineering: dbt, Apache Iceberg, DuckDB, MongoDB, PostgreSQL, SQLite, Redis, Qdrant.

Systems & Infrastructure: Docker*, Git, Kubernetes, Terraform, Linux/Unix, Podman*.

Cloud Services: Amazon Web Service*, Google Cloud Provider.

PROJECTS

Distributed Machine Learning System | Docker, FastAPI, Scikit-Learn, Redis, cuML, Min.io, MLflow, Optuna

- Engineered a heterogeneous MLOps platform enabling one-click training, validation, and deployment across CPU and CUDA-enabled GPU clusters, improving make span and throughput.

Vanilla RAG & Multi-Modal RAG | Celery, Docling, FastAPI, Min.io, PostgreSQL, Qdrant, Redis

- A scalable Multi-Modal RAG system with a dedicated data pipeline, a Polyglot Persistence layer and layout-aware parsing for high-fidelity retrieval of complex unstructured data.

Scikit-Learn in C++ | C++, eigen, vcpkg

- A C++ implementation of Scikit-learn with CUDA and ROCm optimized regression algorithms, achieving 10x speedup over Scikit Learn for large-scale datasets.

Image Search Engine | Docker, PostgreSQL, FastAPI, PyTorch, CUDA, ROCm, uv

- Reducing image search latency by 95% using ROCm and reduce storage usage by 87% using PCA.

Music Recommendation System | Numpy, Pandas, Python, scikit-learn

- Leverages Spotify API data and Non-negative Matrix Factorization (NMF) to generate personalized music recommendations, effectively discovering new tracks aligned with user preferences.

Schedulearn | Docker, FastAPI, Horovod, Python, SQLite

- A lightweight distributed deep learning scheduling system that reduces make span by 50% and increased throughput by 70% across different servers and GPUs.

EDUCATION

National Tsing Hua University

Hsinchu, Taiwan

Master of Science in Computer Science

January 2023

National Tsing Hua University

Hsinchu, Taiwan

Bachelor of Science in Computer Science

January 2021