FADHLILLAH

Backend & AI Engineering for Scaling Businesses
— Remote (international) · Jakarta (WIB · UTC+7)

South Jakarta, Indonesia | +62 851-5704-3131 | fadhlillah949699@gmail.com | WA: wa.me/6285157043131
fadhlillah2.github.io/Bio | linkedin.com/in/fadhlillah2 | github.com/fadhlillah2

WHAT I DO

Software Engineer (6+ yrs) who builds and scales production backend systems and AI/LLM products — currently operating banking microservices serving 2M+ requests/day. Ships fast with AI-augmented engineering — a production SFTP platform delivered in 10 days against a 17-day plan. Experience across healthcare, F&B, and fintech. Open to full-time, remote (international), and consulting engagements.

PROVEN OUTCOMES

- Scale — Spring Boot microservices serving 2M+ requests/day at sub-200ms response times; running 12 microservices handling 500k+ daily transactions (banking)
- Reliability — supports financial systems processing $10M+ daily volume, consistently meeting 48-hour reporting SLAs; sustained 99.5% uptime on a healthcare platform with 100+ APIs
- Speed — a production SFTP platform delivered in 10 days against a 17-day plan, using AI-augmented engineering
- Efficiency — report generation accelerated 700% via query tuning and materialized views; Redis caching cut DB queries by 60% during peak restaurant load (F&B)
- Reach — notification system delivering 50k+ emails and SMS messages daily
- AI — LLM-integrated products end to end: RAG systems, AI chatbots, internal automation; Top 50, Meta Llama Hackathon 2025 (Hacktiv8 Indonesia)

SERVICES

Backend APIs : APIs your business can rely on — design and build of scalable RESTful APIs (Spring Boot, Node.js, Flask)

AI / LLM : RAG systems, AI chatbots, LLM automation — OpenAI, Llama, Groq, LangChain, ChromaDB

Databases : Slow reports fixed at the root — query tuning, indexing, and caching (report generation accelerated 700%)

Microservices : event-driven architecture with Docker, Kubernetes, and Apache Kafka

Consulting : Independent advice before you commit budget — architecture design and technology stack selection for systems at scale

PROOF YOU CAN CHECK

- High-Precision Contract-Advisor RAG — Top 50, Meta Llama Hackathon 2025 (Credential ID 032/H8/META/XI/2025) — github.com/fadhlillah2/llama-docs-auditor
- Rate limiting service in Go, multiple algorithms, distributed deployments — github.com/fadhlillah2/rate-limiter-project-go
- Portfolio, full CV, and LinkedIn — fadhlillah2.github.io/Bio

HOW WE WORK

- Start with a short WhatsApp chat or intro call — wa.me/6285157043131
- Scope, timeline, and code ownership agreed in writing before any work starts
- Build with regular check-ins — code quality held at 90+ (SonarLint, peer review)
- Full handover on completion: source code, documentation, and deployment