

Stefan Baum – Diplom-Informatiker (FH)

Senior Software Engineer • System Architect • DevOps

info@s-baum.dev • +49 162 2906782 • s-baum.dev

Technologies

Backend: Java, Spring Boot, FastAPI, Node.js, Python

Frontend: Angular, TypeScript, NgRx, RxJS

DevOps: Docker, CI/CD Pipelines, Linux, Infrastructure as Code, GPU Server Operations (private)

Databases: PostgreSQL, MySQL, Elasticsearch, Redis, Vector DBs

AI/ML (private): LLM Integration, Local Inference, vLLM, llama.cpp, MLX, LiteLLM, RAG Pipelines, pgvector, Model Evaluation, Document Intelligence, LangGraph, MCP

Tooling: Claude Code, OMP, OpenCode, Pi, Hermes, VS Code, Cursor, Playwright, LangChain, Context Engineering, Coding-agent evaluation

Observability: Grafana, Prometheus, Logging Stacks

Architecture: System Integration, Event-Driven, API Design

Languages

German (native), English

Education

Diplom-Informatiker (FH) — Fulda University of Applied Sciences, Applied Computer Science (1998–2004)

Profile

20+ years of experience in full-stack development, system architecture, and DevOps. All-rounder: distributed systems, legacy modernization, containerization.

Since 2025 as private R&D: local AI infrastructure for open-weight LLM inference and AI-integrated workflows.

Industries so far: enterprise IT, automotive, public sector, e-commerce, travel.

Experience

Since 2019 — Freelance Consultant & Senior Software Developer

- **Embedded Platform** — In-car infotainment for premium automotive; PM, prototyping, multi-gen support (2019–2024)
- **E-Government Digitalization** — Baden-Württemberg, Angular 14, NgRx (2022–2023)
- **Booking Platform Modernization** — Migration of legacy Java services from vendor VM hosting to a containerized Docker architecture; discovery, staging, production cutover, and post-migration stabilization (2025–2026)
- **CMS Migration** (2025–2026)
- **Containerized Environment Setup** — Reproducible dev, staging, and production environments for the platform's services: Docker Compose, Traefik routing, Elasticsearch, CI/CD pipelines, and deployment

runbooks (2025–2026)

- **Dev-Team Consulting** — Knowledge transfer, technical documentation, and ongoing support for the client's development team (2026)
- **E-Commerce Data Integration** — B2B data aggregation, MuleSoft (2020)
- **Local AI Platform** — Private R&D: open-weight LLM inference on NVIDIA Ampere-class GPUs, a dual DGX Spark GB10 pair in tensor-parallel mode, and Apple Silicon; engines vLLM, llama.cpp, MLX; one OpenAI-compatible LiteLLM gateway that routes by privacy class; RAG on PostgreSQL/pgvector; MCP integrations; models and inference engines kept current as they appear (2025–2026)
- **Open-Weight Model Evaluation** — Private R&D: comparison of current open-weight models for development and document-heavy workloads: quantization (NVFP4, GGUF), context length, throughput, and tool-calling reliability; containerized serving monitored with Prometheus and Grafana (2025–2026)
- **Document Intelligence Pipeline** — Private R&D: a staged pipeline that ingests business documents, extracts content with OCR and vision models, classifies them, and files them by taxonomy into a DMS and a RAGFlow index; LangGraph workflow engine (2026)

2017–2019 — Senior Software Developer, Publikat (Signa Sports United)

- B2B data aggregation and e-commerce platform development
- MuleSoft integration and Navision data provider development
- System integrations and ETL pipelines

2015–2017 — Senior Software Developer, Lifestyle-Travel-Team GmbH

- Java-based reservation system for bike tourism
- Full-stack: booking platform, admin, website, tech support

2005–2015 — Software Developer, maxviva Technologies AG

- Full-stack internet booking engine with GDS backend integration
- High-volume transaction platform with 24/7 operations
- Performance optimization and high-availability systems
- Java / Spring