

Stefan Baum – Diplom-Informatiker (FH)

Senior Software Engineer • System Architect • DevOps

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

Profil

20+ Jahre Erfahrung in Full-Stack-Entwicklung, Systemarchitektur und DevOps. Allrounder: verteilte Systeme, Legacy-Modernisierung, Containerisierung.

Seit 2025 private F&E: lokale KI-Infrastruktur für Open-Weight-LLM-Inferenz und KI-integrierte Workflows.

Branchen bisher: Enterprise-IT, Automotive, öffentliche Verwaltung, E-Commerce, Touristik.

Berufserfahrung

Seit 2019 — Freelance Consultant und Senior Software Developer

- **Embedded-Plattform** — Infotainment für Premium-Automotive; PM, Prototyping, Multi-Gen Support (2019–2024)
- **E-Government-Digitalisierung** — Land Baden-Württemberg, Angular 14, NgRx (2022–2023)
- **Buchungsplattform-Modernisierung** — Migration von Legacy-Java-Services von Vendor-VM-Hosting zu einer containerisierten Docker-Architektur; Discovery, Staging, Produktions-Cutover und Post-Migration-Stabilisierung (2025–2026)
- **CMS-Migration** (2025–2026)
- **Containerisierte Umgebungen** — Reproduzierbare Dev-, Staging- und Produktionsumgebungen für die Dienste der Plattform: Docker Compose, Traefik-Routing, Elasticsearch, CI/CD-Pipelines und Deployment-Runbooks (2025–2026)
- **Dev-Team-Consulting** — Wissenstransfer, technische Dokumentation und laufender Support für das Entwicklungsteam des Kunden (2026)
- **E-Commerce-Datenintegration** — B2B-Datenaggregation, MuleSoft (2020)
- **Lokale KI-Plattform** — Private F&E: Open-Weight-LLM-Inferenz auf NVIDIA-GPUs der Ampere-Klasse, einem DGX-Spark-GB10-Paar im Tensor-Parallel-Betrieb und Apple Silicon; Engines vLLM, llama.cpp, MLX; ein OpenAI-kompatibles LiteLLM-Gateway, das nach Datenschutzklasse routet; RAG auf PostgreSQL/pgvector; MCP-Integrationen; Modelle und Inferenz-Engines laufend aktuell gehalten (2025–2026)
- **Open-Weight-Modellevaluation** — Private F&E: Vergleich aktueller Open-Weight-Modelle für Entwicklungs- und dokumentenlastige Workloads: Quantisierung (NVFP4, GGUF), Kontextlänge, Durchsatz und Tool-Calling-Zuverlässigkeit; containerisiertes Serving, überwacht mit Prometheus und Grafana (2025–2026)
- **Dokumenten-Intelligence-Pipeline** — Private F&E: eine mehrstufige Pipeline, die Geschäftsdokumente aufnimmt, Inhalte per OCR und Vision-Modellen extrahiert, sie klassifiziert und nach Taxonomie in ein DMS und einen RAGFlow-Index ablegt; LangGraph-Workflow-Engine (2026)

2017–2019 — Senior Software Entwickler, Publikaat (Signa Sports United)

- B2B-Datenaggregation und E-Commerce-Plattform-Entwicklung
- MuleSoft-Integration und Navision-Datenprovider-Entwicklung
- Systemintegrationen und ETL-Pipelines

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

- Java-basiertes Reservierungssystem für Fahrradtourismus
- Full-Stack: Buchungsplattform, Admin, Website, Tech-Support

2005–2015 — Software Entwickler, maxviva Technologies AG

- Full-Stack Internet-Buchungsmaschine mit GDS-Backend-Integration
- Hochlast-Transaktionsplattform mit 24/7-Betrieb
- Performance-Optimierung und hochverfügbare Systeme
- Java / Spring

Technologien

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

Frontend: Angular, TypeScript, NgRx, RxJS

DevOps: Docker, CI/CD Pipelines, Linux, Infrastructure as Code, GPU-Server-Betrieb (privat)

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

KI/ML (privat): LLM-Integration, Local Inference, vLLM, llama.cpp, MLX, LiteLLM, RAG-Pipelines, pgvector, Modellevaluation, Dokumenten-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

Architektur: System Integration, Event-Driven, API Design

Sprachen

Deutsch (Muttersprache), Englisch

Ausbildung

Diplom-Informatiker (FH) — Hochschule Fulda, Angewandte Informatik (1998–2004)

Sailauf, 7. September 2026