James Staud

Lead Software Architect • AI Platform Engineering • Distributed Systems

SUMMARY

Leader and software architect specializing in production AI/ML platforms, distributed systems, and cloud-native infrastructure. Hands-on experience designing and shipping high-scale AI systems including RAG pipelines, multi-model orchestration, retrieval systems, and Kubernetes-based inference platforms operating under real production load.

Defined SLO frameworks, owned incident response, and drove capacity planning across multi-cluster HA/DR environments, with deep involvement in API architecture, platform reliability, CI/CD automation, observability, and infrastructure-as-code.

Background spans SaaS startups, robotics, logistics, and enterprise AI platforms with strong focus on shipping reliable systems, reducing operational toil, and enabling scalable AI-native engineering practices.

PROFESSIONAL EXPERIENCE

Mouser Electronics — AI Platform Architect / DevOps Architect | 2024–Present

- Personally architected and shipped production AI systems supporting ~85K requests/day (~150 RPS peak) across customer-facing and internal workloads
- Designed and implemented RAG pipelines over 10M+ products and attributes combining structured retrieval, semantic search, embeddings, and LLM reasoning workflows
- Built classifier-driven orchestration pipelines and multi-agent routing systems using Semantic Kernel
- Designed and operated multi-cluster Kubernetes (ARO/OpenShift) platform supporting HA/DR workloads across 2 availability zones and 13 replicasets
- Implemented GitOps-based delivery pipelines (ArgoCD) enabling repeatable, auditable deployment of AI infrastructure and services
- Architected and deployed self-hosted enterprise AI inference platform (vLLM, LLM-d) with integrated development tooling (Jupyter, MLflow) and orchestration on Kubernetes, enabling fully controlled, observable, and scalable AI workflows independent of third-party providers
- Integrated Azure OpenAI services with containerized APIs, retrieval systems, and distributed orchestration layers
- Reduced retrieval latency ~50% through distributed indexing and caching strategies using Snowflake and Azure AI Search
- Built AI-powered document processing pipelines improving throughput ~50x while reducing operational overhead
- Defined SLOs, owned incident response workflows, and established observability standards using Splunk, Dynatrace, Grafana, and Prometheus; drove capacity planning and self-healing automation to maintain platform reliability at scale
- Authored platform architecture documentation, deployment standards, and ADR/RFC-style implementation guidance
- Integrated AI-native engineering workflows using Cursor, GitHub Copilot, and Claude-assisted development patterns
- Adapted architectural direction based on operational telemetry and evolving workload requirements

Cloud 9 Perception — VP Engineering & Co-Founder | 2016–2023

- Personally designed and built distributed AI systems for logistics and manufacturing environments combining computer vision, edge inference, telemetry, and orchestration
- Architected NeuralStack platform for real-time perception, workflow automation, and distributed AI coordination
- Built production ML pipelines for object detection, segmentation, OCR, and edge-optimized inference under strict latency constraints
- Designed event-driven coordination patterns between edge systems and centralized orchestration services

- Developed reliable deployment and telemetry workflows for systems operating in noisy, variable real-world environments
- Integrated hardware, software, and AI systems into production customer environments emphasizing reliability and fault tolerance
- Wrote technical specifications, architecture documentation, and implementation plans spanning robotics, AI pipelines, and infrastructure
- Recruited, mentored, and retained engineers across robotics, computer vision, and distributed systems disciplines; built a culture of technical ownership, peer review, and continuous improvement

Gozova — CTO & Co-Founder | 2015–2017

- Built and shipped cloud-native logistics platform supporting real-time dispatch, routing, and operational workflows
- Designed backend APIs, mobile integrations, and distributed service architecture for multi-market operations
- Established CI/CD and rapid iteration workflows in fast-moving startup environment

EARLY CAREER (2011–2015)

- WRHOWELL — Built real-time control systems, PLC integrations, embedded automation software, and simulation environments for industrial systems
- University of Texas at Arlington — Developed computer vision, gesture recognition, robotics, and HCI research systems; published conference research
- Black Optex — Developed imaging and optoelectronic sensing systems with hardware/software integration
- JFStaud Enterprises — Built early computer vision systems using IR structured light and 3D data processing

TECHNICAL SKILLS

AI/ML Platforms: LLMs, RAG, Semantic Search, Embeddings, Agentic Systems, Multi-Model Orchestration, Prompt Engineering, Semantic Kernel

Cloud & Infrastructure: Kubernetes, OpenShift, Azure, AWS, GCP, Terraform, ArgoCD, GitOps, CI/CD, HA/DR, Multi-Cluster Architectures, API Gateways

Distributed Systems: Event-Driven Architectures, Distributed Retrieval Systems, Microservices, Caching, Observability, Reliability Engineering

Development: Python, JavaScript/TypeScript, SQL, C/C++, FastAPI, Containerization, Buildah, Podman

Observability & Operations: Splunk, Dynatrace, Grafana, Prometheus, Logging, Metrics, Incident Response

KEYWORDS

GenAI, LLM, RAG, AI Agents, Semantic Kernel, Prompt Engineering, Embeddings, Kubernetes, OpenShift, Distributed Systems, Cloud-Native Architecture, Event-Driven Systems, API Gateways, Terraform, ArgoCD, GitOps, CI/CD, HA/DR, Reliability Engineering, SLO, Incident Response, Observability, Splunk, Dynatrace, Grafana, Prometheus, Azure OpenAI, AWS, GCP, Python, FastAPI, Containerization, Multi-Cluster Architectures, vLLM

EDUCATION

B.S. Computer Engineering — University of Texas at Arlington