

Matthew Ezell

Principal Architect | AI Solutions Architecture and Engineering

Birmingham, Alabama matt@thenewguard.ai linkedin.com/in/mattezell github.com/mattezell
thenewguard.ai

Professional Summary

Software architect with 10 years in the role and 15+ years building enterprise software across fintech and enterprise platforms. Most recently co-architect of a natural language query (NLQ) product on an IT spend and asset intelligence platform, where I produced the architectural review, observability and governance designs, a measured prompt-engineering win, and the approved year-long roadmap for the AI practice. Before that, architect and domain lead for an 11-domain compliance reconciliation engine for a large enterprise client, from design through legacy cutover: the system architecture and product design were mine, and a small team came to me for direction across database, API, UI, and ETL. I take AI from demo to production: deterministic compute, model narration, evidence you can audit, and a design gate before every build. I also run The New Guard (thenewguard.ai), a publication for builders shipping with AI, and spend my own time running and tuning local models and building agentic tooling.

Core Skills

Architecture: solution and platform architecture, design gates and ADRs, roadmap authorship, legacy migration and cutover, multi-tenant SaaS, requirements refinement, technical leadership and mentoring.

AI engineering: LLM agents and tool use, Model Context Protocol (MCP), natural language query over governed data, semantic layers and taxonomy grounding, LLM gateways (OpenAI-compatible), eval harnesses and golden sets, prompt experiments with controlled measurement, agent observability (OpenTelemetry GenAI conventions), governance and accountability for agentic workflows, agent coding tooling (Claude Code plugins, GitHub Copilot CLI), local model serving (llama.cpp, vLLM, Ollama).

Engineering: TypeScript and Node.js, JavaScript, C#, Python, PHP, SQL; NestJS, Angular; PostgreSQL, SQL Server, MongoDB, DuckDB; Kubernetes, OpenShift, Helm, ArgoCD, Docker; Jenkins and GitHub Actions CI/CD; REST and SDK design; HCL BigFix REST API; Power BI.

Experience

VS0 LLC, Senior Software Architect, February 2023 to October 2026

Co-architect, natural language query and AI practice (August to September 2026). Brought in by the lead architect after delivering Compliance to co-own the NLQ product on the company's IT spend and asset intelligence platform: an agent that answers finance and operations questions over governed data through MCP tools on a read-only DuckDB materialization, behind an OpenAI-compatible LLM gateway.

- Delivered a full-stack architectural review across six repositories ahead of white-label productization: 32 prioritized recommendations, 26 draft tickets, and a 12-document evidence campaign. Surfaced a broken taxonomy authoring loop, missing observability and evaluation, security and credential-management gaps in a multi-tenant deployment, and an agent-framework migration path (37 imports across 26 files inventoried).
- Designed and ran a controlled prompt experiment (a baseline and 4 variants, 8 questions, 3 runs each, 120 runs). The winning change, injecting the schema's tag vocabulary into the system prompt at boot, shipped and cut tool calls 12 percent and wall clock 11 percent with answer correctness unchanged at 24 of 24 and tool errors down from 4 to 1.
- Authored the approved long-term roadmap: six workstreams, four phases, about 40 milestones each with a design gate and exit criteria, a 13-item decision register with named deciders, and a Jira restructure. Defined the operating model the team now runs on: no implementation ticket without a reviewed design page.
- Wrote the observability design (ten-event structured per-question trace with scrubbing, phased retention from Postgres to an OpenTelemetry collector) and the governance design (workflow by role by evidence accountability matrix, trace access and retention policy, scripted-only gateway administration for customer deployments).
- Authored the LLM gateway ADR (Apache-2.0 OpenAI-compatible gateway, virtual keys per consumer) and a vendor-decoupling posture that keeps the gateway and agent framework swappable behind OpenAI-compatible, MCP, and OpenTelemetry interfaces.
- Produced a 40-feature AI opportunity map and an AI-native platform analysis that defined the trust loop (AI drafts, deterministic validation, human approval, provenance) and the rule that models never compute numbers.
- Designed and stood up a cross-team plugin marketplace for AI coding agents (Claude Code and GitHub Copilot CLI, dual manifests against the Agent Plugins 1.0 spec) to consolidate drifted skills, subagents, hooks, and instruction templates across six repositories, with CI-enforced validation, linting, versioning, and hygiene gates and a 32-skill organization-wide audit.

Solution architect and domain lead, Compliance (August 2024 to July 2026). Architect and subject-matter lead for the compliance layer for a large enterprise client on the same platform, working with a small team: automated reconciliation that keeps thousands of endpoints' maintenance windows, patch and upgrade subscriptions, and security settings continuously in agreement between the compliance database and HCL BigFix, and feeds attestation data to a GRC workflow tool and an audit system. The system architecture and product design were mine. I designed the initial database schema (no DBA for the first half of the project) and the initial API surface, then directed the developers assigned to each while the UI engineer owned the front end.

- Replaced a 21-module legacy platform (three NestJS APIs, six Angular frontends, eight shared libraries) after a full-stack gap analysis across API, database, ETL, and UI.
- Designed 21 processors across 11 domains as request, monitoring, and discrepancy triads (async submit, poll, reconcile) with environment gating, dry run, error budgets, and freshness watermarks, then led the cutover from the legacy system and the hypercare window.
- Solved eventual-consistency and stale-data races by splitting monitoring into a separate daily review stage so compliance was never evaluated against a not-yet-refreshed nightly ETL result.

- Moved security-settings sync to BigFix from XML computer-setting payloads to flat-file routing to clear a character-limit ceiling.
- Replaced a legacy SQL Server decommission job, which had once sat disabled for six weeks unnoticed, with a scheduled executor that retires hosts in BigFix from a queue of Ansible-submitted requests (dry-run by default) plus a watermark-based health monitor that catches a stalled executor in hours.
- Wrote the complete documentation set (42 documents, about 114,000 words: 11 domains, architecture, 12 decision records, every page citing the source code it describes) in the weeks after cutover, prioritized by an audit of 101 real support and developer questions clustered into 15 themes.

Solution architect, large client account (2023 to 2025). Owned requirements refinement and solution design across the account. Led the migration from bare-metal Kubernetes to OpenShift, reworking the Jenkins pipelines, container builds, and Helm deployment for 16 services across two production data centers. Modernized every product codebase, including the rebuild of the reporting module as an Angular application over a SQL Server data layer and the move of its reports API from the legacy BI solution to Power BI, and advised other groups in the client organization. Mentored VSO and client engineers.

Fiserv, Enterprise Senior Software Engineer / Architect Advisory II, 2012 to 2022

Fintech and banking platforms. Analyzed complex business requirements and proposed solutions on leading industry technologies; drove technology stack decisions; architected and engineered SDKs for internal and client development teams; led client training and demo sessions; performed code reviews and direct client consulting; documented technical systems; served as subject matter expert on emerging technologies and approaches; mentored and provided directional consulting to multiple organizations within the company.

Whiteboard IT, Associate Developer / Technical Consultant, 2010 to 2012

Delivered n-tier solutions for clients across business domains on then-emerging technologies (Node.js and NoSQL when both were contested choices). Advised clients on solution options and cost, maintained and enhanced legacy solutions, administered vSphere, and built custom features for a large medical company's public-facing SharePoint instances.

Independent Work

The New Guard (thenewguard.ai), founder and writer, February 2026 to present. An AI publication for builders who ship. Hands-on deep dives on getting agentic tools working, including what broke and why. Ongoing experimentation with local model serving, quantization, tool calling, and agent workflows on personal hardware.

Education and Certifications

Bachelor of Applied Computer Science, Troy University, 2008, cum laude (3.548 GPA)

CompTIA A+, 2008