

Jonathan Kosar

Senior Software Engineer — AI Platform

Nashville, TN / jonathankosar@gmail.com / (615) 862-2605 / dakosar.com / in/jonathankosar / gh/jonathanmkosar

Senior engineer on the AI Platform team, owning the company's semantic-search product end-to-end — runtime, ingestion, LLM-provider strategy, and the CI/CD behind it. Builds for **scale-to-zero economics** — true serverless that pays for work, not idle capacity — and ships through an **agent-driven, AI-first workflow**, treating AI as the default way to automate from development through operations. Drives both as patterns for other teams to adopt.

EXPERIENCE

Senior Software Engineer — AI Platform

Oct 2024 — Present

HealthStream · AI Platform

- Own the company's semantic-search product end-to-end — now the leading **candidate to replace the legacy search system**; onboarded **7+ enterprise customers** across **21 production releases with zero P1/P2 outages**.
- Cut cost per request **~85%** as search volume grew **42x**, architecting the platform to **idle at zero cost** — pay for work performed, not standing capacity — anchored by a purpose-built embedding cache.
- Own **LLM-provider strategy**: embedding APIs carry **no latency or uptime SLA**, so architected multi-region hedging and fallback to deliver the enterprise search latency and availability providers don't — plus a major model upgrade.
- Built an **agentic development harness** — reusable agent skills and commands automating CI/CD, now used by **two other teams** — sustaining **1,172 commits across 17 repos** this cycle as solo product owner.
- Cut content propagation from the legacy source from **~24 hours to ~30 minutes** with an **event-driven** webhook ingestion service, alongside a rebuilt nightly path.
- Rebuilt the production **vector search** (embedding + cosine-similarity) to **enterprise-scale capacity** — each query re-architected to run serverless on sharded Lambda with a faster Go runtime, holding search latency low under load.
- Operationalized and optimized **semantic compression** — an LLM preprocessing pass that cleans noisy learning content and normalizes embeddings end-to-end, improving the consistency and accuracy of cosine-similarity search.

Senior Software Engineer

Oct 2014 — Oct 2024

Bank of America · Trading & Risk · first 2 yrs via FDM · Platinum awards 2017-18

- Shipped a capacity-management & simulation app delivering **\$30M/yr in savings (2023)** by reclaiming idle capacity; built a hardware-provisioning workflow adopted as a business-unit standard and a batch-job monitoring service running **9+ years** in production.
- Led and mentored engineering teams of **3–10+** across time zones; set TDD and design standards through weekly knowledge-sharing, raising code quality across critical financial systems.
- Refactored a single-threaded poller in a real-time **Tornado WebSocket** service into a non-blocking, multi-threaded model, clearing heavy-I/O bottlenecks across **7 backend processes**.
- Rebuilt a real-time data pipeline on bidirectional, compressed WebSockets with event-driven React — cutting load times from **minutes to seconds** and client memory by hundreds of MB.
- Automated large-scale **Python migrations** (2.7 → 3.10) with scripts, regex, and unit tests, cutting migration effort **~4x**.
- Drove a multi-organization **RHEL upgrade across 12K hosts in 3 global regions** and shipped a Jinja-based low-code approval/notification system that became the org's primary maintenance-notification tool.

Computer Network Engineer

Dec 2004 — Dec 2008

U.S. Air Force

- Built and maintained **network, VoIP, and cryptographic infrastructure** spanning analog and digital systems; sustained **\$80M+** in equipment and led peer groups through technical training.
- Awarded the **Meritorious Service Medal** for outstanding leadership and technical performance in a critical deployed environment during Operation Iraqi Freedom; held Top Secret clearance.

SKILLS

CORE STACK

Python · Cursor/Claude Code · React · Serverless

AI & AGENTS

Agentic development (skills, harnesses) · MCP servers · Semantic search at scale · Embeddings · LLM-provider strategy (multi-region) · RAG · Model migration & rollout

CLOUD & INFRA

Event-driven serverless on AWS (Lambda · DynamoDB · S3 · API Gateway) · Terraform IaC · CI/CD automation · Multi-region resilience · Data pipelines

PLATFORM

Operational agentic automation & observability · Dependency auditing · Security review & remediation

SELECTED PROJECTS

Serverless AI Search

Embedding + cosine-similarity vector search re-architected to run per query on Lambda — sharded for throughput — powering production AI search at enterprise scale.

Agentic dev harness

Reusable agent skills and commands that automate CI/CD end-to-end — agentic tooling built to make one engineer operate like a team.

Semantic compression

An LLM pass that cleans raw learning content and normalizes embeddings end-to-end — sharper, more consistent cosine-similarity matches.

Keystone PERSONAL

Purpose-built SQLite "stones" (finance, routines, todos, fantasy sports), each exposed to LLM agents via a dedicated MCP server. One domain → one SQLite → one MCP.

EDUCATION

B.A. Computer Science

Hunter College, CUNY — New York
2009 — 2014

A.S. Electronic Systems Technology

Community College of the Air Force
2004 — 2005