

Siddharth Deshpande

SOFTWARE ARCHITECT | AI SYSTEMS | ENGINEERING LEADER

+91 78478 96052 · dsiddharth2@gmail.com
siddharthdeshpande.com · github.com/dsiddharth2
linkedin.com/in/dsiddharth2

12 years of turning ambiguous problems into production systems – ERPs, AI platforms, multi-agent pipelines – across startups and enterprise. I don't just write code; **I own the outcome end to end**. At **Apra Labs** I built the AI and data platform layer serving **300+ US facilities**: **7 AI agents** and **5 production AI systems**, a GraphRAG engine over **150K documents**, conversational AI across a **4 TB database**, and a FinOps agent that cut cloud spend **30%** – while hiring and mentoring **20+ engineers** and growing with the company **from 10 to 55**. Before that, a logistics ERP built from scratch to **2,500 daily users**, **70K dockets/month** and **100K hits a day**, at **MildlyClassic**, where our team also built the software that got **Wifi Dabba into YC W17**. Separately, I co-founded **UXgage**, a behavioral analytics startup **₹5L funded by Deshpande Startups, Hubli**.

SYSTEMS I'VE ARCHITECTED AND SHIPPED

GraphRAG Knowledge Engine

Enterprise knowledge graph pipeline

- Python Azure Durable Functions Cosmos DB FAISS gpt-4o-mini
- Builds **knowledge graphs** with **entities, relationships, topics and communities** via **vector similarity**
 - Two-stage ingest**: **seconds-to-searchable indexing**, then **graph enrichment** in the background
 - 90K entities** and **150K+ relationships** across **11 Cosmos collections**
 - Outcome** — **Support knowledge reusable at scale**; duplicate ticket resolution dramatically reduced

Multi-Agent Conversational AI

Supervisor-pattern agent orchestration

- Python LangChain LangGraph Azure SQL Azure SignalR React
- Supervisor pattern routing to **5 specialized agents** across **~30 registry-driven tools**
 - Natural-language SQL with **permission rewrite**, seven-collection RAG, **100+ report types** over a **4 TB database**, SignalR streaming of the reasoning trace
 - Outcome** — Report generation reduced from **days to seconds**; non-technical staff querying data independently

Cloud Cost Intelligence Agent

Automated FinOps analysis

- TypeScript Azure Cost Management Resource Graph Claude Sonnet
- Nightly FinOps agent**: **19 collectors** across 3 phases, **13 anomaly detectors**, 4-tier breach thresholds
 - 20+ standing questions** answered deterministically; the LLM handles only reasoning, with a **doer-reviewer loop** checking claims against raw files
 - Outcome** — **30% cloud cost reduction**; prioritized recommendations delivered nightly, teams now own their spend

Infrastructure Performance Agent

Health scoring & regression detection

- TypeScript Azure Monitor SQL DMVs Query Store Service Bus
- Composite **0-100 health score** per service, with **rolling 30-day baselines** for **regression detection**
 - Azure Monitor metrics and SQL DMV telemetry sampled at **15-second intervals**, **13 anomaly checks**, 4-tier breach thresholds
 - Outcome** — Degradation surfaced **before user impact**; alert fatigue significantly reduced

Secure Logistics ERP

PAN India logistics platform

- PHP AWS React Node.js MySQL
- End-to-end ERP built from scratch** — Finance, Operations, Booking, Hub Management, CRM, Route Planning
 - 2,500 daily users**, **70K dockets/month**, **100K daily hits**, zero downtime through scaling
 - Outcome** — **End-to-end shipment visibility** for all clients; manual processes reduced by over 80%

AI Code Review Pipeline

CI/CD-integrated automated reviews

- Python Azure DevOps GPT REST APIs
- Reviews pull requests with **inline comments and star ratings**. **10+ languages**, **penalty-based scoring**, **fix verification mode**
 - Outcome** — **Review bottleneck for senior engineers eliminated**; consistent scoring across all repos

Search Index Sync Engine

Real-time & bulk search indexing

- C#/.NET 8 Azure AI Search Service Bus Durable Functions
- Bulk indexing** via Durable Functions and **real-time incremental sync** via Service Bus, with **index versioning** and a **dead letter queue**
 - Outcome** — **Entity search across 7 people-data entity types** made near-instant; search freshness under **5 seconds**, zero-downtime schema migrations

UXgage – Behavioral Analytics

Enterprise usability platform

- PHP Neo4j Cassandra JavaScript
- Captures clicks, mouse movements, text input and URL transitions to **reconstruct actual user workflows**. **Built the entire stack from scratch**
 - Outcome** — **₹5L seed funding**; idea to funded product in under 6 months

EXPERIENCE

Apra Labs

Associate Architect · previously Senior Software Engineer
Aug 2019 – Present

Built the **AI and data platform layer** for a unified facility management product serving **300+ US facilities**. Managing **3 clients**, mentoring a **team of 4**, **architected all systems end-to-end**. **Campus hiring lead for 4 years**; grew with the company **from 10 to 55**.

MildlyClassic

Engineering Team Lead
May 2016 – May 2019

Built and scaled a **logistics ERP from scratch** to **2,500 daily users** handling **70K dockets/month** for PAN India secure jewellery logistics. **Hired 7 engineers**, established a **4-week release cycle**, migrated **DigitalOcean → AWS in 3 months**.

Sister company · Wifi Dabba → YC W17: built **byte-counting and usage sync software** on Wi-Fi hardware, and the **Google NBU analytics platform**.

UXgage

Co-Founder & CTO
Oct 2015 – Apr 2016

Took an idea **from zero to a shipped product in 6 months** — won a **₹5L cash prize** from Deshpande Startups, Hubli, with office space and Azure infrastructure granted separately.

Earlier

2010 – 2016

Samvit Learning, Founder — software training institute founded during MCA; trained **250+ students**. **Firaa.me**, side project at IdeaFlask — travel blogging platform, **500+ organic users**. **dotCORD IT Solutions**, Intern — **3 years** at GIT Incubation Center.

OPEN SOURCE

Plug — package manager for Claude Code skills; install and manage reusable Skills, Commands and Agents from GitHub-hosted registries. **14 packages**, on npm as *plugvault*.

CodeHawk — AI pull request review pipeline. Runs in CI, scores code quality and posts inline comments. **16 language rule sets**, **394 unit tests**.

PHP ZXing — PHP wrapper for the ZXing barcode library. **277K+ Packagist installs**, **133 GitHub stars**, featured on the upstream ZXing repo.

MooPin — Pinterest-style column grid for MooTools, incl. infinite scroll. **~120 lines**.

TECHNICAL EXPERTISE

Languages & Frameworks — Python · TypeScript · C# · .NET 8 · PHP · React · Node.js · LangChain · LangGraph

AI & Data — GPT / LLM integration · FAISS · Vector search · GraphRAG · Multi-agent systems

Infrastructure — Azure (Durable Functions, Service Bus, Cosmos DB) · AWS · Neo4j · Cassandra

Practices — System design · CI/CD · DevOps · FinOps · Hiring & mentorship

EDUCATION & RECOGNITION

Master of Computer Applications (MCA) — Gogte Institute of Technology, Belgaum

Leadership Excellence, Apra Labs (Nov 2024) · **Customer Appreciation**, Apra Labs (Apr 2023) · **Guiding Vision**, Pravega X Fest