Swapnil Surdi

San Jose, California
+1 (213) 294 8993 • swapnil.surdi@icloud.com • surdi • swapnilsurdi
Google Scholar - 66E21KUAAAAJ

Summary

Senior Software Engineer with 7+ years of experience specializing in full-stack development, API architecture, and AI-driven solutions. Proven track record in 0-to-1 environments, moving from monoliths to scalable microservices and building agentic RAG systems. Expert at translating complex customer requirements into high-leverage tools, including custom Model Context Protocol (MCP) integrations. Passionate about API-first design, automated workflows, and technical consulting.

Skills

Languages: Python, TypeScript/JavaScript, Go
Backend & APIs: FastAPI, Node.js, Django, GraphQL, REST, Kafka, RabbitMQ
LLM & AI Systems: Model Context Protocol (MCP), LangChain, RAG Pipelines, Agentic Workflows, Vector DBs
Cloud & Infrastructure: AWS (ECS, Fargate, Lambda, DynamoDB, S3), Docker, Kubernetes
Data & Caching: PostgreSQL, MongoDB, Redis, DynamoDB
Open Source: Published MCP tools, MCP contributions, active in AI/LLM communities

Experience

Senior Software Engineer Sept 2021 – July 2025
Treatment Technologies & Insights Inc Los Angeles, CA

- Architected production agentic RAG system using FastAPI and MCP tools for documentation/compliance workflows, enabling multi-hop reasoning across codebases, Confluence, and Jira with 768-dimensional embeddings in Qdrant, accelerating documentation retrieval by 10x and streamlining regulatory compliance verification for ISO 13485 certification
- Built custom MCP servers in Python for healthcare data access, integrating with LangChain orchestration layer to enable LLM agents to query patient health data, submit questionnaires, and analyze survey responses through authenticated FastAPI endpoints with HIPAA-compliant audit logging
- Architected 0-to-1 HIPAA-compliant microservices platform on AWS Fargate supporting 4 enterprise healthcare clients, designing core services (Profile, Health Data, Organizations, Trials) with standardized error handling, distributed tracing, and auto-scaling achieving 99.9% uptime
- Reduced patient data API latency from 12s to sub-second (P95) through DynamoDB partition key redesign, batch processing, and Redis caching at API Gateway layer for 1000+ patient datasets
- Built AWS S3/CloudFront CDN architecture for medical imaging (OHIF viewer), reducing 500MB study load times from 90+ seconds to under 25 seconds through worker-based processing and edge caching
- Designed cross-platform dynamic survey engine with version-controlled JSON schemas, enabling new client onboarding in days instead of months while supporting 12+ languages including RTL, with unified rendering across React, iOS, and Android
- Implemented distributed error tracking with partial UUID correlation across frontend and backend, reducing MTTR by 70% through centralized watchdog service implemented in FastAPI
- Developed Go-based testing framework enabling parallel end-to-end testing and rate limiting validation, achieving 85% test coverage
- Mentored three junior engineers on system design, MCP development, and AWS best practices, conducting code reviews and establishing team standards that improved code quality and accelerated feature delivery

Application Developer Oct 2016 – June 2018
Apprely Technologies Pune, India

- Built financial platform with Kafka/RabbitMQ for asynchronous processing, integrating QuickBooks, Zoho, and Plaid/Stripe payment gateways
- Developed Android applications with Django REST backends for healthcare appointment management and GPS-based outdoor navigation

Associate Software Engineer Oct 2015 – Oct 2016
Accenture Pune, India

- Automated Salesforce regression testing for Virgin Media UK using Selenium, processing 60K+ pricing tests and reducing 2 months manual QA to under 1 hour. Awarded Ace Prize.

Selected Projects

MCP-Cache @hapus/mcp-cache

- Published NPM utility for MCP agents optimizing LLM token usage through intelligent caching of large tool outputs, reducing token costs by up to 80% for repetitive queries. Used in production by multiple AI applications.

MCP-Confluence-Jira (Healthcare Compliance) Production Internal Tool

- Built FastMCP server enabling LLM agents to search and query ISO 13485 compliance documentation across Confluence/Jira, powering agentic workflows for regulatory document generation and CAPA tracking. Reduced documentation retrieval time from minutes to seconds.

MCP-HealthData (HIPAA-Compliant) Production Internal Tool

- Developed FastAPI-based MCP server with OAuth2 authentication enabling LLMs to securely access patient health records, submit questionnaires, and analyze survey data through structured tool definitions. Implemented comprehensive audit logging for HIPAA compliance.

Launchlab swapnilsurdi/launchlab

- Built Docker-based homelab orchestration with one-liner deployment for self-hosted services (Jellyfin, Immich, Paperless-ngx).

Education

Master of Science, Astronautical Engineering 2020
University of Southern California GPA: 3.5
Bachelor of Engineering, Electronics & Telecommunications 2015
Savitribai Phule Pune University GPA: 3.5
Certifications: Six Sigma Green Belt (USC Marshall), HIPAA Compliance, GDPR Data Security

Publications

- **Publication:** *Space situational awareness through blockchain technology*, Journal of Space Safety Engineering, 2020.