

AKASH RAMASANI

San Francisco, CA 94107 | +1-510-935-6846 | akashramasani2705@gmail.com
github.com/akash-ramasani | linkedin.com/in/akash-ramasani

PROFILE

Software engineer with 5+ years of experience shipping **production backend systems, cloud infrastructure, and LLM-powered products end-to-end**. Currently the sole engineer owning a production platform rebuild in TypeScript, spanning the API, Temporal workers, MCP server, Next.js applications, CI/CD, and Stripe billing. Builds daily with AI-assisted development tools (Claude Code, Cursor) and has shipped production LLM and agentic features on the Anthropic and OpenAI APIs.

SKILLS

Languages: TypeScript, Python, JavaScript, SQL, Java

Backend / APIs: Node.js, Fastify, FastAPI, Flask, Spring Boot, REST APIs, WebSockets, Temporal, Model Context Protocol (MCP)

AI / LLM: LLM and AI agent development, Anthropic API, OpenAI API, custom MCP servers, tool use, prompt engineering, prompt caching, structured outputs, AI-assisted development (Claude Code, Cursor)

Cloud / DevOps: AWS (App Runner, ECS, ECR, S3), Docker, Kubernetes, Terraform, GitHub Actions, GitHub OIDC, CI/CD, Linux

Databases / Data: PostgreSQL, Firestore, Drizzle ORM, Prisma ORM, Hibernate / JPA, Ignite caching, schema design, zero-downtime migrations, ETL pipelines

Observability / Testing: Prometheus, Grafana, Vitest, pytest, JMeter, Locust, production incident response, log-based debugging

Frontend / Payments: React, Next.js (App Router), Tailwind CSS, Stripe (subscriptions, metered billing, webhooks)

PROFESSIONAL EXPERIENCE

Runn Technologies

Jun 2026 – Present

Senior Software Engineer

Remote

- **Sole engineer shipping a production TypeScript platform rebuild end-to-end:** 8 services spanning a Fastify API with 200+ routes, a Temporal worker running 14 workflows, an MCP server, a headless page renderer, and Next.js customer, admin, and marketing apps, backed by 1,300+ automated tests.
- Runs in production for **1,200+ customers**, with 3,000+ ad campaigns launched and \$500K+ in ad spend managed through the platform.
- Built an **LLM pipeline on the Anthropic and OpenAI APIs** that reads a customer's website, drafts schema-validated ad copy, and launches a Google Ads campaign through an idempotent Temporal workflow that resumes safely on retry.
- Cut monthly **AWS spend by 52%** through an infrastructure audit and service rightsizing, with 65% projected after legacy infrastructure retirement.
- Designed a **79-table PostgreSQL schema with 71 versioned migrations** (Drizzle ORM), and applied every production migration with zero data loss.
- Integrated **16 external providers**, including Google Ads, Meta Marketing, Microsoft Ads, TikTok Ads, Stripe, OpenAI, and Anthropic, and built Stripe billing end-to-end with subscriptions, metered usage reporting, webhooks, coupons, and partner pricing.
- Replaced Auth0 with a **first-party session-based authentication** system that revokes access on the next request after sign-out, suspension, or password change.
- Automated deployment of 6 services with **GitHub Actions, GitHub OIDC, Docker, Amazon ECR, and AWS App Runner**, with no long-lived AWS credentials.

Compass Group

Aug 2024 – May 2026

Senior Software Engineer

San Jose, California

- Shipped **LLM-powered automation** on the OpenAI API with MCP-based tooling, partnering with product and engineering to replace manual business processes across the organization.
- Led a **zero-downtime DB2 to PostgreSQL migration** with end-to-end data validation, improving query performance and data-access efficiency.
- Designed high-volume **REST APIs in Spring Boot and JPA** handling millions of daily transactions, with Ignite and Hibernate caching to reduce database load and response times.
- Modernized Spring Batch jobs and migrated them to **AWS ECS**, cutting batch processing time by 30%.
- Built Python dashboards for **real-time cluster monitoring** and internal reporting, speeding up incident debugging.
- Integrated OIDC authentication, ran JMeter performance tests, and owned **production support**, resolving critical incidents and documenting system architecture.

Oracle

Software Development Engineer II

May 2022 – Jul 2024

Hyderabad, India

- Engineered scalable backend services and internal APIs in **Python, FastAPI, and Flask** on Oracle Cloud Infrastructure, improving request latency by 30%.
- Refactored legacy modules into containerized **microservices**, reducing deployment time from hours to minutes.
- Automated deployment and monitoring with **Docker, Kubernetes, Terraform, and Jenkins**, reducing manual infrastructure effort by 40%.
- Improved reliability and observability with **Prometheus and Grafana**, maintaining 99.9% uptime and debugging production incidents from logs and metrics.
- Implemented **OAuth2 and JWT authentication** and ran Locust benchmarks that improved throughput by 25%.

Indian Council of Agricultural Research

Software Development Engineer I

Jan 2021 – May 2022

Hyderabad, India

- Engineered a reporting platform used in **Indian Parliament reporting workflows and by 5,000+ scientists**, cutting report-generation time from hours to seconds through optimized SQL and ETL pipelines, role-based access control, dynamic charts, and PDF exports (Flask, REST APIs).
- Designed computer vision and geospatial **machine learning pipelines** for rice-crop classification from satellite imagery, reaching 96 to 97% accuracy with a TensorFlow CNN pipeline served through Flask endpoints.

SELECTED PROJECTS

JobWatch: Open-Source AI Job-Tracking Platform

github.com/akash-ramasani/job-watch

- Built and open-sourced an end-to-end job-tracking platform (React 19, Firebase Gen 2) that **monitors 950+ companies across nine ATS platforms** (Greenhouse, Ashby, Workday, Oracle Recruiting Cloud, Eightfold, and more) using smart pagination, Firestore TTL, and asynchronous processing.
- Integrated a **Claude-powered AI agent and a custom MCP server** driving automated job scoring and personalized cover-letter generation; cut LLM latency and API cost by 80%+ with Anthropic prompt caching.
- Designed **hybrid job scoring**: a rule-based score for every posting with LLM evaluation reserved for high-scoring matches, and cut Firestore reads by 99% with aggregation documents and caching.
- Built a browser extension and background worker that **auto-fill job applications with AI** on Greenhouse and Ashby, reducing manual application effort by 95%.

RapidRxNow Operations Hub

- Architected an internal catalog-management application for pharmacy operations with **Next.js App Router, Prisma ORM, and PostgreSQL**, including RBAC, custom Edge Middleware, and NextAuth authentication.
- Engineered **data-ingestion workflows** for POS datasets with Server-Sent Events for background progress, batch processing, and multi-store synchronization.

3-Node Kubernetes Homelab and Distributed Log Monitoring

- Built a three-node **Kubernetes cluster** on repurposed hardware with a FastAPI service that exposes Linux logs through REST APIs and scans multi-GB log files with reverse chunked reads, covered by pytest.

EDUCATION

California State University, East Bay

Master of Science, Computer Science

Hayward, California

KL University

Bachelor of Technology, Computer Science and Artificial Intelligence

Hyderabad, India