Anshul Tayal

Principal AI Full-Stack Cloud Engineer

EXECUTIVE SUMMARY

Principal-level AI Full-Stack Engineer with **10+ years of experience** designing and scaling **cloud-native, AI-driven platforms** used by **millions of users daily**. Proven leader in architecting **high-throughput distributed systems**, **production AI/ML pipelines**, and **real-time customer-facing platforms** on Google Cloud.

Recognized for driving **multi-million-dollar revenue impact**, **processing 300K+ orders/day and 3M+ events/day**, and delivering **$500K+ annual cost savings** through architectural optimization. Adept at **leading 30+ engineer teams**, mentoring senior engineers, and translating ambiguous business problems into **scalable, resilient, AI-powered solutions**.

AI-driven professional with hands-on experience in Vertex AI, Gemini Agents, LLM integration, conversational AI, and predictive modeling systems. Leverages generative AI platforms, including ChatGPT, Google Gemini, GitHub Copilot, Perplexity, NotebookLM, and Claude, to automate workflows, accelerate software development, enhance research synthesis, and improve data-driven decision-making. Proven ability to operationalize AI/ML solutions into daily business processes to drive efficiency, scalability, and innovation.

CORE TECHNICAL LEADERSHIP

- **AI/ML Systems:** Vertex AI, Gemini Agents, LLM Integration, Conversational AI, Prediction Systems
- **GEN AI Tools:** ChatGPT, Gemini Code Assists, Github Copilot, Perplexity, NotebookLM, Claude, Codex
- **Backend & Full Stack:** Java, Spring Boot, REST APIs, Microservices
- **Cloud & Platform:** Google Cloud (GKE, Pub/Sub, Maps API), Kubernetes, Docker
- **DevOps & IaC:** Terraform, Jenkins, Harness, CI/CD, Observability & Monitoring
- **Architecture:** Distributed Systems, Event-Driven Design, High-Availability Platforms
- **Leadership:** Technical Strategy, Staff Mentorship, Architecture Governance

AI Innovation & Architecture Portfolio

AI System Design Copilot — AI-Powered Architecture & Distributed Systems Assistant

- Designed and developed an AI-powered system design copilot that converts product ideas into scalable distributed system architectures, HLD summaries, API contracts, database strategies, trade-off analysis, and deployment recommendations.
- Built a Spring Boot and Next.js full-stack architecture with provider-selectable GenAI workflows, supporting local demo generation plus OpenAI and Vertex AI integrations through a normalized architecture response contract.
- Implemented prompt-aware architecture generation for real-world use cases, including delivery tracking, ride-sharing, chat systems, media platforms, recommendation/search systems, analytics products, SaaS workspaces, IoT telemetry, and payments.
- Engineered structured outputs for services, APIs, data stores, event flows, Mermaid diagrams, scalability plans, failure modes, cost optimizations, and exportable Markdown engineering design docs.
- Added PostgreSQL-backed saved design history and workspace-scoped records as a foundation for authenticated collaboration and long-lived architecture review workflows.
- Built an interactive React dashboard with prompt-driven generation, provider selection, Mermaid architecture visualization, saved history loading, and Markdown case study export.

AI-Powered Cloud Cost Optimization & FinOps Advisor

- Architected and developed an AI-powered cloud cost optimization platform that analyzes infrastructure utilization, billing trends, and workload patterns to generate actionable savings recommendations across multi-cloud environments.
- Designed scalable microservices-based backend using Java, Spring Boot, REST APIs, and PostgreSQL to process cloud billing datasets and infrastructure telemetry in near real time.
- Built intelligent recommendation engines for identifying idle resources, over-provisioned compute instances, unattached storage volumes, and inefficient autoscaling configurations.
- Developed AI-driven rightsizing models to recommend optimized CPU, memory, and storage configurations, reducing projected cloud infrastructure costs by 30–50%.
- Implemented event-driven data ingestion pipelines for cloud billing exports, utilization metrics, and infrastructure inventory using asynchronous processing patterns.
- Created interactive dashboards using Next.js, Tailwind CSS, and modern visualization libraries to display spend analytics, anomaly detection, optimization opportunities, and savings forecasts.
- Integrated AI-generated operational insights and remediation summaries to help engineering teams prioritize high-impact cost optimization initiatives.
- Designed extensible cloud provider abstraction layers to support future integrations with Google Cloud, AWS, and Azure billing APIs.
- Containerized the application stack using Docker and Docker Compose, enabling simplified local development, CI/CD readiness, and cloud-native deployment workflows.
- Authored production-grade architecture documentation, API contracts, onboarding guides, and demo scripts to support developer adoption and stakeholder presentations.
- Implemented automated cost anomaly detection logic capable of identifying unexpected infrastructure spend spikes and underutilized resources.
- Built portfolio-grade full-stack architecture showcasing expertise in distributed systems, cloud engineering, AI-assisted operations, and FinOps automation.

PROFESSIONAL EXPERIENCE

Karmuu Technologies
Technical Advisor & Architect — AI-Powered Job Search Platform
(Summer 2026 Internship Project) -
https://karmuutech.com/f/summer-2026-internship---ai-powered-super-intelligent-job-search

- Architected an AI-powered intelligent job search platform enabling personalized job discovery, resume matching, and recruiter-candidate optimization using LLMs and vector-based search.
- Designed scalable microservices architecture using Java, Spring Boot, REST APIs, and cloud-native deployment patterns for high-availability recruitment workflows.
- Led the end-to-end technical vision for an AI-driven hiring ecosystem integrating semantic search, resume parsing, profile enrichment, and intelligent recommendation engines.
- Defined system architecture for real-time candidate-job matching using embeddings, Retrieval-Augmented Generation (RAG), and AI ranking pipelines.
- Advised on integration strategy for external hiring and enrichment APIs, including Greenhouse, Lever, Workable, Proxycurl, and job aggregation providers.
- Designed AI workflows enabling recruiters to automatically shortlist candidates based on skills, experience relevance, behavioral fit, and semantic profile similarity.
- Built architecture roadmap for scalable multi-tenant SaaS deployment supporting employers, recruiters, and candidates across web and mobile platforms.
- Created backend service contracts, database entity models, event-driven workflows, and API gateway strategies for enterprise-grade extensibility.
- Guided implementation of AI-assisted resume optimization, interview preparation assistance, and personalized job recommendation features.
- Evaluated build-vs-buy decisions for external AI and recruitment integrations to optimize engineering velocity, operational cost, and platform scalability.
- Collaborated on product strategy, MVP prioritization, and investor-focused technical roadmap for an AI-first recruiting platform.
- Defined scalable cloud infrastructure strategy leveraging Kubernetes, Docker, CI/CD pipelines, observability, and infrastructure-as-code practices.
- Proposed AI governance and model monitoring strategies, including feedback loops, ranking quality evaluation, recommendation explainability, and retraining pipelines.

- Mentored engineering contributors on distributed systems design, API standards, scalable backend architecture, and AI integration best practices.
- Designed recruiter analytics dashboards and AI-powered insights systems to improve hiring efficiency, candidate engagement, and conversion metrics.

HCL America Inc.—Principal Full-Stack Cloud & AI Engineer

Louisville, KY | Jan 2022 – Present

Principal technical owner for AI-powered digital commerce, omnichannel retail, logistics, and POS modernization platforms at enterprise scale.

Impact Highlights

- Architected distributed systems processing 300K+ orders/day and 3M+ events/day with sub-second latency and enterprise-grade reliability
- Designed AI-powered Quote Time Algorithm (QTA), achieving >90% delivery-time prediction accuracy
- Reduced Google Maps API costs by 50%+, saving ~$500K annually through architectural optimization and governance initiatives
- Led 30+ engineers across backend, cloud, DevOps, AI/ML, eCommerce modernization, and omnichannel initiatives
- Modernized legacy commerce and POS ecosystems into scalable cloud-native microservices architectures supporting omnichannel customer experiences
- Integrated generative AI tools into engineering workflows to improve collaboration, documentation quality, productivity, and solution scalability
- Utilized Google Gemini and AI-assisted development tools to accelerate delivery cycles, reduce coding defects, and improve engineering efficiency

Key Contributions

- Designed and delivered AI-driven prediction systems using Vertex AI, powering real-time delivery estimation and intelligent logistics workflows
- Owned end-to-end platform architecture for real-time delivery, omnichannel commerce, POS modernization, prediction, and tracking systems, balancing latency SLAs, resiliency, scalability, and cloud cost optimization at enterprise scale
- Led enterprise eCommerce modernization initiatives, transforming legacy monolithic platforms into scalable event-driven microservices architectures using Java, Spring Boot, GKE, Pub/Sub, Docker, and Kubernetes
- Architected omnichannel integration platforms connecting mobile applications, web ordering systems, in-store POS, loyalty systems, payment gateways, and third-party delivery marketplaces into a unified commerce ecosystem
- Designed and productionized the AI Quote Time Algorithm (QTA), achieving >90% delivery-time accuracy while balancing model precision, sub-second inference latency, and infrastructure efficiency
- Spearheaded POS modernization strategy by introducing API-first integration layers, decoupled store services, real-time synchronization patterns, and scalable backend integration frameworks
- Authored and governed Architecture Decision Records (ADRs) covering event schemas, distributed system patterns, AI inference workflows, integration standards, and cloud cost tradeoffs, establishing architectural governance across teams
- Established event-driven architectures on Google Cloud Platform (Pub/Sub, GKE, Vertex AI), enabling resilient real-time processing, fault isolation, and horizontal scalability for high-throughput enterprise workloads
- Designed omnichannel order orchestration and real-time event processing systems supporting seamless customer experiences across web, mobile, delivery, and in-store channels
- Led architectural strategy for PapaTrack, Drone Delivery Experience, omnichannel integrations, and next-generation commerce initiatives, aligning business, operations, and engineering stakeholders on long-term modernization roadmaps

- Integrated third-party logistics providers, delivery aggregators, payment vendors, loyalty systems, and external commerce platforms at enterprise scale
- Improved customer engagement through real-time order visibility, synchronized promotions, delivery tracking, loyalty integration, and personalized commerce experiences across digital and physical touchpoints
- Drove cloud cost governance initiatives, reducing Google Maps API spend by 50%+ (~$500K annually) through caching strategies, API optimization, traffic controls, and vendor contract improvements
- Built and governed CI/CD pipelines, Infrastructure-as-Code frameworks (Terraform/Jenkins), observability standards, and deployment automation practices to improve operational reliability and engineering velocity
- Established engineering standards, architecture governance processes, coding guidelines, and scalable integration frameworks adopted across multiple engineering teams
- Optimized platform reliability through centralized monitoring, distributed tracing, proactive alerting, incident response frameworks, and performance tuning initiatives
- Mentored senior and staff engineers through architecture reviews, technical strategy sessions, and engineering coaching, influencing technical excellence across 30+ engineers
- Served as senior technical escalation owner for production incidents, scalability bottlenecks, AI model anomalies, and distributed system failures, driving root-cause analysis and long-term corrective actions
- Partnered with product, operations, infrastructure, and executive leadership teams to define modernization roadmaps, scalability strategies, and future-state enterprise commerce architectures

HCL America Inc.—Architect Owner

Aug 2021 – Jan 2022

- Led and owned the technical delivery of PapaTrack, supporting 300K+ daily orders by coordinating architecture, development, and production readiness across multiple engineering teams
- Architected and launched Drive-Up Pick-Up during COVID, rapidly expanding fulfillment capabilities while operating under accelerated timelines and heightened reliability constraints
- Improved team delivery velocity by ~40% by restructuring sprint planning, backlog prioritization, and cross-team dependency management
- Defined cloud-native architecture blueprints and reference patterns adopted across programs, standardizing platform design and accelerating delivery of new initiatives

HCL Technologies—Lead Engineer

Noida, India | Jul 2018 – Aug 2021

- Designed and implemented scalable microservices platforms on Google Cloud using Kubernetes and Pub/Sub, addressing throughput, reliability, and service isolation requirements
- Delivered 50%+ API cost reduction by implementing intelligent Google Maps API usage controls, caching strategies, and request optimization
- Led hands-on development and system design while establishing code quality standards through structured code reviews, refactoring initiatives, and best-practice enforcement
- Mentored engineers transitioning from mid-level to senior roles through design guidance, code reviews, and ownership of complex, business-critical features

HCL Technologies—Software Engineer

Dec 2015 – Jul 2018

- Built and maintained enterprise backend systems using Java, Python, and REST APIs, supporting core business workflows and internal services
- Delivered end-to-end features across design, development, testing, and deployment, owning tasks from requirements clarification through production release

- Improved application reliability through performance tuning, refactoring legacy components, and resolving production stability issues

EDUCATION

- **PGDM – Business Analytics** — IMT Ghaziabad
- **PG Diploma – Advanced Computing**—C-DAC Bangalore
- **B.Tech – Computer Science** — Jaypee University of Information Technology

Certification

- **Building with the Claude API - Anthropic Certification**
- **Claude 101 - Anthropic Certification**
- **Claude Code 101 - Anthropic Certification**
- **Introduction to Claude Cowork - Anthropic Certification**
- **Introduction to Model Context Protocol - Anthropic Certification**
- **Introduction to Agent Skills - Anthropic Certification**
- **Introduction to Subagents - Anthropic Certification**
- **Model Context Protocol: Advanced Topics - Anthropic Certification**
- **Claude with Google Cloud's Vertex AI - Anthropic Certification**
- **NVIDIA AI Technical Curriculum (2025)**
- **NVIDIA AI Sales Curriculum (2025)**
- **Machine Learning & Artificial Intelligence Certification** - Programming Hub - 1705694546779
- **Securing and Integrating Components of your Application** - Coursera - 7LKFV6CWD6J4
- **App Deployment, Debugging, and Performance** - Coursera - BJJ3Z7YNMD2S
- **Getting Started With Application Development** - Coursera - ZYBB8ENBPPWU
- **Google Cloud Associate Cloud Engineer Exam** - Coursera - X6568NAHSMNM
- **Google Cloud Platform Fundamentals: Core Infrastructure** - Coursera - YJ88LHN96N4V
- **Developing Applications with Google Cloud Platform Specialization** - Coursera - S937RAFFJPDA
- **Java Programming Language using Java SE6** - Oracle Workforce Development
- **Fundamentals Certification Course** - Programming Hub - 36466250
- **AMCAT Certified Data Processing Specialist** - Aspiring Minds - 1524513-211
- **AMCAT Certified Collections Specialist** - Aspiring Minds - 1524513-197
- **Python Basic** - HackerRank - 3279048cbff5
- **Python Intermediate** - HackerRank - ad7a12ce619a
- **Python Advanced** - HackerRank - 85749e630eb9

AWARDS & RECOGNITION

- **O2–O5 League of Extraordinary Awards**—HCL Technologies (2020–2023)
- **AI/ML Hackathon Winner 1st Place**—Providing Optimal Tip—2023
- **2 Good & 3 Hive Certificates of Appreciation**—HCL (2024–2026)
- **Top Cheese Award**—HCL/Papa John’s Program