

Dhiraj Kumar Sharma

Seattle, WA | dhirajksharma922@gmail.com | linkedin.com/in/imdhirajsharma

Professional Summary

Senior Software Engineer at Microsoft AI with 10+ years shipping production systems end to end, 7 of them at Amazon. Architected Armada, a multi-agent orchestration platform delivered across GitHub Copilot, Copilot CLI, and VS Code — parallel agent fan-out, single-owner resource scoping, result reconciliation, and verification gates before completion — including the VS Code extension UI and MCP-based observability tooling in TypeScript. Also lead the migration of Bing Ads' multi-region Hadoop/Spark data platform to Azure, owning the networking layer, the secrets migration, and a per-team cutover path. Previously owned tier-1 services at ~6M events/min sustaining 99.99%+ availability, built the Alexa in-app purchase product from zero to ~\$4M annual revenue, and led the Hey Disney partner integration.

Experience

Microsoft | Senior Software Engineer, Microsoft AI — Bing Ads, Data Platform as a Service (Jun 2026 – Present)

- Architected and built Armada, a multi-agent orchestration platform that decomposes complex engineering work, dispatches independent tasks to parallel AI sub-agents, protects shared resources through single-owner scoping, reconciles results, and gates completion on verification.
- Built the Armada VS Code extension in TypeScript — @armada chat participant, fleet tree visualization, agent and task metadata, status tooltips, language-model tooling, and the Fleet Hub dashboard — plus the Copilot CLI plugin (/armada command, armada-fan-out skill, MCP-based fleet observability), kept consistent across all three surfaces by a canonical core with generated adapters so the integrations cannot drift.
- Established resource-aware concurrency, dependency management, lifecycle policies, verification gates, and model right-sizing so agent fleets run reliably and cost-consciously; designed local-first fleet-state processing with sanitized summaries, path redaction, and data minimization.
- Shipped Copilot CLI plugins used across the org: /ado-pr for governed Azure DevOps pull-request inspection, approval, and merge — two-phase approval, short-lived action tokens, policy checks, guarded bypass — and /ground-rule for persistent AI operating standards backed by a canonical JSON store with atomic writes and protected markers.
- Built a reusable Terraform/CDKTF platform toolkit of 29 skills spanning security baselines, provider governance, state migration and drift resolution, module scaffolding, cost estimation, environment promotion, and PR quality gates.
- Lead the migration of Data Platform as a Service (DPaaS) — the multi-region Hadoop/HDFS and Spark platform where all Bing Ads bidding data lands — from on-prem datacenters to Azure, and built its Azure networking layer from the ground up including Private Link to consuming services.
- Designed and implemented the HashiCorp Vault to Azure Key Vault migration: event-driven key and secret sync, a reconciliation path to catch drift between stores, staged deprecation, and a client-by-client cutover onto AKV.
- Designed the AKS job-execution layer around ephemeral VMs — the scheduler provisions a dedicated VM per job and tears it down on completion, isolating tenant workloads and removing idle capacity.
- Own the deployment motion for each team moving onto DPaaS: discovery, technical scoping, design review, hands-on migration of their Spark/ETL workloads, and production cutover support.

Amazon | Software Development Engineer (Apr 2019 – Apr 2026)

- Primary owner of tier-1, high-throughput push-notification services processing ~6M events/min on AWS (SQS/SNS/DynamoDB/ECS/Fargate/Lambda, CDK), sustaining 99.99%+ availability (99.999% during peak readiness) with defined SLIs/SLOs across multi-team dependencies.
- Designed event-driven architectures with back-pressure, retries/DLQs, idempotency, and fault isolation; improved tail latency and reduced operational load through scalable data-access patterns.
- Owned ORRs and on-call operations (incident response, postmortems/RCA); defined SLIs/SLOs and alerting standards, created reusable service/CDK templates and runbooks, and improved monitoring & alerting, load testing, and game days for revenue-critical services.
- Built Alexa Mobile In-App Purchase (IAP) and subscription backend, delivering ~\$4M annual revenue (2023) across ~100K subscriptions; implemented entitlements, billing integrations, and resilient purchase processing.
- Led backend design and launch readiness for partner integrations (e.g., Hey Disney purchasing) and media portability; reduced account-linking load time by ~85%.
- Revamped multi-room media control platform using event-driven microservices, enabling distributed state synchronization and increasing user conversion by 45%.
- Optimized AWS Rekognition Facet fleet through CPU/memory profiling and tuning, delivering ~\$14M YoY reduction in annual infrastructure spend.
- Amazon Security Certifier (since 2020); led cross-team design reviews and mentored engineers, creating reusable CDK/service templates adopted by multiple teams to raise the bar on security and production readiness.

Oracle | Senior Software Engineer (Apr 2017 – Apr 2019)

- Built backend microservices and REST APIs in Java/Spring Boot for enterprise transaction workflows, improving reliability and maintainability of core services.
- Reduced production errors by ~45% by profiling JVM services (VisualVM), tuning garbage collection, and optimizing database queries and indexes.
- Partnered with product and QA to harden releases (test automation, defect triage, and rollout checklists), improving deployment quality and reducing customer-reported issues.
- Provided technical leadership for a 10+ engineer team: design reviews, code reviews, and performance debugging; helped drive improvements in service scalability and operational stability.

Care.com | Senior Software Engineer (Nov 2016 – Mar 2017)

- Improved acquisition funnels by optimizing cross-browser landing pages and email templates, increasing click-through rates by ~25% and reducing bounce rates by ~15%.
- Implemented instrumentation and event-logging libraries to improve analytics quality and enable A/B testing, contributing to a ~21% increase in conversion.

Samsung Semiconductors | Senior Software Engineer (Jul 2016 – Nov 2016)

- Built marketplace features for Artik modules/products to improve discoverability and purchasing workflows, contributing to ~27% revenue growth.
- Streamlined partner support/communication portals for the Artik cloud ecosystem, reducing response times by ~35% through workflow and UX improvements.
- Enhanced features and analytics on Ruby on Rails services and web surfaces, increasing engagement by ~43% and reducing bounce rate by ~25%.

Selected Projects

- **Armada — independent productization:** Productized Armada into a standalone, portable agent-orchestration framework with canonical orchestration policies, three Copilot surface adapters, architecture documentation, and installation workflows — adoptable across repositories without replacing the host model or assistant.
- **Beam — full-stack Android TV launcher on abandoned hardware:** Reverse-engineered a discontinued Meta Portal TV with no docs, no root, and no app store, then shipped a working product solo: the complete 10-foot TV UI (universal search, QR sign-in, D-pad navigation) over a channel-ingestion and app-integration backend serving 700+ live channels in 11 languages and 25+ sideloaded apps. In daily use as the device's primary interface.
- **Job-search automation layer:** Automation built on the open-source career-ops pipeline (santifer/career-ops, MIT): a scheduled board scanner that screens postings against a role rubric, a resume generator that edits the master .docx in place so fonts, spacing, and styles stay byte-identical rather than being re-derived, and a submission flow that always stops for human review before anything is sent.

Skills

- **LLM Systems (built):** Multi-agent orchestration and reconciliation, parallel agent fan-out with resource scoping, verification gates, MCP servers and clients, tool/function calling, context budgeting, model right-sizing, privacy-preserving state handling
- **AI Developer Platforms:** GitHub Copilot extensibility, Copilot CLI plugins, custom agents and skills, VS Code Extension API, Claude Code; developer experience and self-service tooling
- **Data & Distributed Processing:** Hadoop/HDFS, Spark, ETL and scheduled aggregation pipelines, multi-region cluster operations, on-prem to cloud migration; Redis, RabbitMQ, Azure Service Bus
- **Backend / Distributed Systems:** System Design, Distributed Systems, Microservices, Event-Driven Architecture, Scalability, Fault Tolerance, high availability, Latency/Throughput, Capacity Planning
- **Languages / Frameworks / APIs:** TypeScript, JavaScript/Node.js, Python, Java/JVM, Kotlin, SQL; React, React Native; Spring Boot; REST, GraphQL; OAuth, AuthN/AuthZ
- **Cloud / Containers / IaC:** Azure (AKS, Key Vault, Private Link, Service Bus, Azure DevOps), Terraform, CDKTF, Bicep; AWS (ECS/Fargate, Lambda, EC2, S3, DynamoDB, SQS/SNS, CloudWatch, CDK); Docker; Kubernetes
- **Observability & Operations:** Monitoring & alerting, Dashboards, On-call, Incident response, Postmortems/RCA, SLI/SLO, Runbooks, ORRs
- **Engineering:** CI/CD, Jenkins, Maven/Gradle, SonarQube; JUnit/Mockito/TestNG; Code reviews; Profiling/Debugging; Agile/Scrum/Kanban

Certifications

- Amazon Security Certifier (2020 – 2026) — design-review authority for production security readiness
- Red Hat Certified Engineer (RHCE); Docker & Kubernetes

Education

M.S., Computer Science — Texas A&M University–Kingsville

B.Tech, Computer Engineering — Rajasthan Technical University