# VIJAY PATEL

New York, NY | (908) 801-3451 | vijpatel7@gmail.com

## SOFTWARE DEVELOPMENT ENGINEER – BACKEND

Backend engineer who ships high impact products fast. Specializes in scalable distributed systems with proven ability to identify high-value problems, communicate vision, and combine system design with product thinking to architect end-to-end solutions: reduced operational costs by $7.7M+ annually, improved process productivity by 90%, and generated $25M+ in incremental profit. Shipped a full-stack web app from zero to 2k+ visitors in 1 month by leveraging AI to accelerate delivery.

## EXPERIENCE

**Amazon** **New York, NY**
*Software Development Engineer II* *July 2021–Present*

**Project Geese: Distributed Cache Migration**

- Built a high-performance 99.99% availability distributed writer-reader cache to retrieve campaign data handling 1M+ TPS
- Created multi-threaded cache client with request sharding that performs parallel cache lookups and data processing to achieve fault tolerance in case of thread failures achieving end to end cache data retrieval and processing in <3ms P99, 75% reduction
- Migrated from on host caches with seldom stale kafka streams that reduced campaign over delivery by $7.7M

**Fenix Priming: Ad Serving Priming Strategy**

- Drove a 4-team initiative (led 5 engineers) building a cross-service priming signal that pre-computes ad sourcing, reducing real-time workflow P99 latency by 20%, improving coverage by 3% with fewer timeouts, and driving $25M+ in annual profit
- Led design for and implemented a performant cache with optimized P99 put and get operation latency from 80ms -> 15ms
- Identified bottlenecks due to payload size and serialization time, implemented Protobufs to reduce payload by 90% to <200kb
- Expanded scope by leveraging latency savings to increase ad auction ad density by 1.5x, leading to 1.2% higher cost-per-click

**Shazam - SB Ad Diagnostic Tool**

- Built an AI-powered diagnostics platform integrating automated database queries, live traffic replay, and LLM analysis to troubleshoot ad campaign issues, reducing ticket resolution time by 83% (30 to 5 minutes) and incoming tickets by 90%
- Drove cross-team collaboration with 3 partner teams to expand tool coverage, enabling API access and data integration
- Navigated ambiguity by directly engaging with support team to understand pain points and iterating on solution through weekly feedback including decreasing overall tool runtime by 66% (12 to 4 minutes) by sharding database queries

## SELF-INITIATED PROJECTS

**Temple's 25th Anniversary Website - njrajatmahotsav.com (nights & weekends)**

- Independently built and launched a full-stack web application in <6 weeks, leveraging AI to accelerate development and rapidly master React/Next.js, delivering a production-grade web app that attracted 2k+ visitors in the first month post-launch
- Designed a mobile-friendly, scalable, low-latency architecture hosted on Vercel integrating animation libraries (GSAP, Framer Motion), Cloudflare for media optimization, Supabase for real-time database operations, Resend for email delivery
- Owned product iteration by gathering user feedback, re-prioritizing features, and shipping improvements in <24 hours

**Database Query Agent – Built for Amazon SB Ad Serving**

- Identified recurring pain point with data analysis and took initiative to prototype an LLM-based agent to generate Athena queries from natural language requests, and forced adoption across 5 teams through demos, wiki, and management visibility
- Implemented a RAG architecture that categorized user requests, retrieved relevant template queries, and synthesized final Athena queries with schema context and LLM-generated field mappings, reducing overall query generation time by 80%
- Engineered a multi-shot reasoning pipeline with automated field validation that decomposed complex analytical requests into sequential LLM calls for table selection, field identification, and query assembly, then verified all fields existed in target tables

## Skills

- **Languages & Frameworks:** Java, Python, TypeScript, React, Next.js, CSS, Tailwind, Spring Boot, Node.js, Protobufs
- **Backend:** Distributed Systems, Redis, Memcached, High Availability, High Performance, Microservices, System Design
- **AWS**: EC2, ECS, NLB, S3, VPC, Elasticache, CloudFormation, IAM, Lambda, Bedrock, Event Bridge, CloudWatch
- **AI/ML:** LLMs (Claude), MCP, RAG Architecture, Prompt/Context Engineering, GenAI, Vector Databases

## EDUCATION

**Georgia Institute of Technology -** *M.S in Computer Science, Specialization in Machine Learning* *Jan 2024 –Present*
**University of Michigan – Ann Arbor -** *B.S. in Computer Science* *Sept 2017–2021*