Renadh Chowdhury

646-220-6996 | renadhc@gmail.com | renadh12.github.io | github.com/renadh12

Summary

Staff Platform Engineer with 8+ years, currently building production grade ML inference systems for real-time fraud detection and risk scoring at Visa AI Platform. Expert in ML Infrastructure, Network, Kubernetes, Distributed Systems, ML Lifecycle solutions and Developer Productivity.

Technical Skills

Languages: Python, Rust, Java, TypeScript, JavaScript, Groovy, SQL
Frameworks: FastAPI, Uvicorn, Node.js, Flask, Serverless, Google Cloud SDK
Network & Infra: F5 Load Balancer, DNS, Istio, VXLAN, Calico CNI, Kubernetes, cert-manager
Libraries: kube-rs, tokio, k8s-openapi, PySpark, ONNX Runtime, pandas, NumPy, Matplotlib, Serverless Offline, Ray Serve, KubeRay
Data: Apache Kafka, Redis, Spark
Cloud: AWS (Lambda, S3, EKS), Azure (AKS, Key Vault, Managed Identities, HDInsight Kafka), GCP (GKE, Compute Engine, Google CAS)
Platform: Docker, Prometheus, Grafana, Jenkins, GitHub Actions, Packer, Helm

Professional Experience

Staff Machine Learning Engineer, Real-Time Inference & AI Platform Feb. 2024 – Present
Visa Inc. *Austin, TX*

- Led technical strategy and ground-up development of 2 production grade AI/ML platforms (GCP GKE, Visa MKE on-prem) powering real-time fraud detection, serving 13+ models including CyberSource ($623B payment volume, 6.4B transactions), Visa Direct, and Visa A2A Payments, handling a combined total compute of 12k+ transactions per second.
- Successfully migrated 9 CyberSource models from on-prem bare metal platform to on-prem Kubernetes platform for global transactions, reducing resource allocation from 600 CPU to 80 CPU while maintaining availability and reducing latency from 50ms to 20ms (99.95%), with extensive optimizations implemented in every layer - infra, network, k8s, application.
- Achieved 99.95th percentile latency of 4.99ms at 500 TPS on Google Cloud Platform with only 36 vCPU/50GB for one CyberSource model deployment in Dammam, Saudi Arabia while meeting data localization compliance requirements.
- Contributed to core ML inference serving framework (Python/Rust/FastAPI/Uvicorn) coupled with advanced Kubernetes patterns implementation including multi-container pods (1 proxy: N inference containers), leveraging both horizontal and vertical scaling, achieving 40% better resource efficiency than traditional sidecars.
- Redesigned k8s network architecture through extensive research and benchmarking: network layer → Istio Gateway/Virtual Service/DestinationRule → K8s Service → Pod → Container, optimized for high-throughput, low-latency inference without service mesh overhead.
- Bootstrapped platforms with full enterprise stack: data streaming (Kafka, Redis, Hadoop, Flink), observability (Prometheus, Grafana, Humio, Kibana), certificate and vault secret management, PAN/non-PAN encryption, inference log pipelines for Data Science and Research teams while maintaining enterprise compliance (ITDR).
- Established end-to-end ML lifecycle framework with reusable infrastructure sizing blueprint (latency, throughput, traffic patterns, stateful vs stateless, feature count) and automated CI/CD pipelines (Jenkins, Packer, Helm)—reducing time to market from 6 months to 14 weeks.
- Led strategic Mirantis cluster optimization achieving 53% namespace, 46% pod, 50% CPU reduction, enabling migration from bare metal to on-prem Kubernetes instead of costly vendor solutions.
- Mentored 12 engineers while collaborating with 10+ cross-organizational teams (Operations & Infra, Model Engineering, Cybersecurity, Data Science); led Production Reliability Engineering knowledge transfer for multi-datacenter deployments across platforms.

Senior Software Engineer, Risk & Authentication Sept 2022 – Feb 2024
Visa Inc. *Austin, TX*

- Designed Java-based Karate framework for real-time API validations across 5 data centers for EMV 3D-Secure Authentication flows with Kafka/Vault integration in Production.

Software Engineer, Developer Productivity Dec. 2020 – Sept. 2022
FOX Corp. *New York, NY*

- Built event-driven CI/CD platform with AWS Lambda/SNS/SQS and TypeScript/Python microservices for developer productivity.

Founding Partner, Lead Engineer Feb. 2020 – Sept. 2022
Next Level Sports Management *New York, NY*

- Co-founded and architected platform serving 100+ athletes and universities globally, demonstrating 0-to-1 product development.

Software Engineer Mar. 2019 – Aug. 2019
Red Ventures *Austin, TX*

- Led test automation for TPG US ($50M revenue, 10M visitors/month) using Cypress/Selenium, improving reliability by 10%.

Software Engineer Aug. 2017 – Mar. 2019
Affinity Solutions Inc *New York, NY*

- Developed Python test automation for financial platform (10M users), reducing bugs by 15%.

Personal Projects

PlatML | *Rust, Axum, React, GKE* | github.com/renadh12/platml 2025

- PoC ML platform with sub-minute model deployment via Rust-powered APIs and model management

KubeRust - Rust based K8 Controller | *kube-rs, k8s-openaip, tokio, kind, docker* Dec 2024 – Present

- PoC Kubernetes controller for dynamic node scaling based on pod utilization metrics

projectX | *PySpark, Docker, Azure* | github.com/renadh12/projectX 2023

- PoC End-to-end ML pipeline with feature engineering and model evaluation on AKS

Education

Western Governor's University Jan 2025 - Present
Bachelor of Science in Software Engineering *Online, USA*

NYC Data Science Academy Oct 2019 - Dec 2019
Data Science with Python: Machine Learning, Data Analysis *New York, NY*