Yi Shen

+1 (608)334-8858 | yishen.org@gmail.com | www.linkedin.com/in/yi-shen

Professional Summary

Senior Software Development Engineer with 8 years at Amazon, architecting **Tier-1 real-time ML platforms** handling **petabytes of data & millions of QPS** with 99.99% availability. Organizational leader driving **cross-company innovation** by influencing strategic technology roadmaps to unlock new capabilities and **60% cost reductions**. Strong in **AI automation**, **distributed systems**, and delivering high-impact solutions with **strong ownership** and engineering excellence.

Professional Skills

- **Languages:** Java, Python, TypeScript, SQL, Shell
- **Cloud & Infra:** AWS, Kubernetes, Docker, Flink, Git
- **AI & Data:** LLM, RAG, MCP, Spark, Hadoop, EMR
- **System:** Serverless, Microservices, Event-Driven, Caching

Work Experience

Amazon.com Services LLC *Seattle, WA*

- **Senior SDE (SDE 3)** July 2023 - Present
  - **Architect** of a Tier-1 **real-time** self-service ML feature platform serving hundreds of internal scientists, handling **petabyte-scale** data and **millions of queries per second (QPS)** with 7 millisecond retrieval latency and 99.99% availability.
  - Designed and led a team to implement a centralized **Serverless Control Plane** (using AWS API Gateway, Lambda, Step Functions, etc.), reducing infrastructure provisioning time by **90%** and enabling full platform automation.
  - Led strategic evaluation of distributed databases (**TiDB, Aerospike, FoundationDB, etc.**) and emerging compute frameworks **(Ray, Pulsar)** via deep-dive POCs and **cross-company collaboration**, influencing product roadmaps to unlock enhanced system capabilities and achieving **60%** cost savings.
  - Developed an internal AI assistant using **RAG, MCP** and **Strands Agent**, speeding up scientist onboarding, authoring and end-to-end feature development workflows **from 2 weeks to less than 2 days**.
  - Optimized platform retrieval performance by implementing batch parallelization and thread pool tuning with profiler, reducing p99 latency from **900ms to 40ms**, saving Amazon **$50M+** in fraud losses annually.
  - Optimized high-throughput batch ingestion system performance, supporting **10+ Terabytes of data per day** and further accelerating ML feature time-to-market for fraud detection.
  - Served as a **Security Reviewer** for cross-team initiatives, auditing designs across the organization to ensure compliance with strict security standards and mitigating critical risks for **Kubernetes clusters** and **AWS infrastructure**.
  - Championed engineering excellence **across the organization** by mentoring engineers and defining best practices for large-scale distributed systems.
- **SDE 2** March 2020 - June 2023
  - Implemented real-time automatic hotkey and anomaly detection as mitigation mechanisms to defend against **bot attacks**.
  - Pioneered automatic cross account **Infrastructure as Code (IaC)** provisioning systems deploying resources efficiently, reducing manual onboarding work by **90%**.
  - Reduced metadata load latency from **30s to 4s** through cache optimizations.
  - Revamped platform UI to significantly improve user experience and productivity for the self-service feature workflows.
  - Engineered robust streaming data pipelines with **AWS Kinesis**, enabling real-time fraud detection with second-level latency.
- **SDE 1** June 2018 - March 2020
  - Contributed to building and maintaining the internal model training and inference platform for fraud detection and abuse prevention, serving hundreds of internal scientists.
  - Developed a critical feature simulation framework that reduced feature authoring and testing iteration time from **weeks to days** by enabling offline simulation without impacting production, accelerating the deployment of new fraud detection models.

Education

- **University of Wisconsin-Madison** Madison, WI
  B.S. Computer Engineering, double major in B.S. Computer Sciences ***(3.94 / 4.0 GPA)*** *May 2018*
- **Stanford University (Summer Session)** Palo Alto, CA
  Intensive Studies In Computer Science Certificate ***(4.2 / 4.3 GPA)*** *Summer 2016*