HARSH P. BAJAJ

Redmond, WA | (206) 825 0909

harshpbajaj@yahoo.co.in | https://github.com/harsh543 | https://www.linkedin.com/in/harshbajaj543

I am an Engineer with 7+ years of experience designing and implementing scalable, high-performance, and distributed systems across cloud and AI infrastructures. Passionate about pushing the boundaries of LLM and VLM performance optimization across modern GPU architectures. Proven track record in building telemetry-aware platforms and fault-tolerant AI workflows at Microsoft and AWS, leveraging cloud-native tooling, GPU acceleration, and predictive analytics for real-world generative AI applications.

Work Authorization: H1B visa until September 2027

EXPERIENCE

Microsoft-Redmond, WA **July 2021 - Present** *Software Engineer II (Silicon, Cloud Hardware Infrastructure*

● **Architected a resilient telemetry pipeline using Spark, Delta Lake, and Synapse**● Built a failure prediction system using classification and forecasting models on GPU telemetry data; deployed models to AI agents running on Azure Service Fabric, enabling real-time inference and intelligent fault response across distributed hardware.

● **Led cross-functional collaboration across hardware, ML, and platform teams** to implement scalable inference workflows with precision tuning, telemetry-aware adjustments, and fault-tolerant triggers.

● ***Developed failure prediction models for NVIDIA H100/H200 and AMD MI300 systems****, integrating real-time inference into AI agents deployed on Azure Service Fabric.*● Developed a system to process Doorbell events from Hardware in Ring 0 system in distributed system environment in Kubernetes cluster and docker to collect GPU hardware telemetry using C++ and store them in Kusto for hardware diagnostics.

● **Reduced end-to-end inference response time by optimizing data flow and container orchestration**, enabling scalable deployment across thousands of virtual machines.

Amazon Web Services-Seattle, WA **Jul 2019 - July 2021** *Software Development Engineer(AWS Identity)*

● Developed a passwordless sign on system for AWS SSO used OTP based Multi factor authentication in Java.

● Worked on launching the TOTP(time-based one-time password) functionality for AWS SSO Console using Java. Leveraging the IAM access policies stored in MongoDB to determine authentication mechanism and authorization of access.

● Built a shadow Machine learning model infrastructure and a complete AI system design using AWS

FireEye-Milpitas, CA **May 2018 - Aug 2018** *Software Engineering Intern(Email & Cloud Security)*

- Developed a distributed system to fetch loglines from Email security servers that were used to scan and filter malicious emails and integrated it with Jira. Then made a feature engineered dataset for K-Means **Clustering algorithm** to cluster the text data as per FireEye's log format for **Email security threats**.

EDUCATION

Master of Science in Computer Science May 2019 University of Illinois, Chicago

Relevant Coursework: Cloud Computing, Machine Learning, Distributed Systems, Artificial Intelligence Safety, Building Secure Computer Systems, Big Data, Deep Learning

B.Tech Computer Science and Engineering May 2014 Vellore Institute of Technology**-**Vellore, India

Relevant Coursework: Object-Oriented Programming, Data Mining, Linear Algebra, Statistics and Probability, Graph Theory, Computer Architecture and Organization, Computer Networks

SKILLS

Programming Languages: Python, Java, C, C# , C++, PHP, JavaScript, Go

Libraries/Framework: TensorFlow, Keras, Kubernetes, Docker, Express, Splunk, Flask, AWS EMR, AWS Kinesis, Node, Redfish, Synapse, ML Flow, .Net, Kusto explorer, SFMC explorer, Sci-kit learn, Apache Spark, Git, ARM, REST API

Platforms: Azure Service Fabric, Azure functions, Azure ML, Azure Storage account, Google Cloud ML Engine, AWS CloudFormation, Azure AI foundry

PROJECTS

- Developed an end-to-end RAG system that indexed custom documents using VectorStoreIndex and retrieved contextually relevant data for LLM-based responses.
- Developed a Generative AI system to predict stock market based on the data from various datasources like Finance websites, News, and Federal reserve and deployed the LLM on Bedrock. Then used Generative to predict the direction of stock market and target ticker price.
- Built a scalable RAG system using Azure AI Search and OpenAI models to generate grounded responses over enterprise data, enabling semantic and hybrid search with secure, production-grade deployment. Integrated prompt templating, streaming capabilities, and evaluation workflows to improve response relevance and factual accuracy.