Harsh P. Bajaj

Redmond, WA harshpbajaj@yahoo.co.in
(206) 825 0909 www.linkedin.com/in/harshbajaj543 https://github.com/harsh543

Senior Software Engineer

Software Engineer with experience building scalable, secure, and intelligent enterprise applications for Global Fortune 500 technology companies, ensuring high availability and performance. Proven expertise in full-stack development, AI/ML integration, cloud infrastructure, and microservices. Passionate about creating tools that enhance workforce productivity and collaboration and mentoring junior developers in ML and design patterns.

Skills

Full-stack development | Scalable distributed systems | SaaS platforms | Web application architecture
Cloud infrastructure | AI/ML | GenAI | Microservices | RESTful APIs | CI/CD | DevOps | BFF Patterns | OOD/OOP
Python | Java | C++ | C# | JavaScript | Go | Node.js | ReactJS | AngularJS
Spring Boot | .NET | TensorFlow | Keras | Synapse | MLFlow | Apache Spark | PyTorch | MCP
Azure (Service Fabric, ML, Functions, Compute) | AWS (Sagemaker, EC2, CloudFormation) | Docker | Kubernetes | Podman | App Insights | Geneva Monitoring
Kafka | Kusto Explorer | Git | Jira | Visual Studio | JetBrains
Windows | Linux | Android | iOS
SQL | NoSQL | DynamoDB | MongoDB | CosmosDB | OAuth | SSO | Authentication & Authorization
LLMs | Agents | Web Services | Container orchestration | Design patterns | Testing frameworks

Professional Experience

Microsoft - Redmond, WA July 2021 - present
Software Engineer II (Silicon, Cloud Hardware Infrastructure)

- Designed and implemented a scalable orchestration platform using Databricks, Apache Spark, and Azure Synapse, enabling seamless data processing across global enterprise datasets. Reducing the latency by upto 2 seconds.
- Developed a machine learning pipeline using Python and Synapse ML to predict GPU failures from telemetry data, deploying models as AI agents in Azure Foundry. Predicting proactively for about 13 days ahead.
- Built a distributed telemetry ingestion system using Docker, Kafka, C++, and .NET, integrating with system-level APIs and secured via Entra ID and certificates. Reducing latency by 5% and availability improvement to 99.99%.
- Collaborated cross-functionally with hardware and software teams to deliver high-performance, secure, and maintainable internal tools for cloud infrastructure monitoring.
- Developed a scalable distributed system in docker environment using Pilotfish to fetch hardware telemetry on Rack level of GPU SKUs.
- telemetry using DSTS and calling system level APIs in C++ and C# .NET to collect telemetry from Kafka and store in Kusto using certificates and managed entra-id.
- Made a monitoring application in Grafana to metrics from telemetry collection App to ensure availability and reliability of system increase by 25%.

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

- Engineered TOTP-based authentication for AWS SSO Console using Java, enhancing secure access for enterprise users. Increasing customer usage by 200% on secure platform.
- Built and maintained CI/CD pipelines using AWS CodeDeploy and CloudFormation, improving deployment efficiency and reliability. Reducing deployment time by 75%.
- Designed a shadow AI system using AWS Sagemaker, EC2, and CloudWatch, integrating Jupyter notebooks and Java-based services for internal analytics. Decreasing latency per request by 5 seconds.

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

- Developed a custom K-Means clustering algorithm for logline analysis, integrated with Jira, automating issue tracking and improving internal support workflows.

- Made an `alert system` using SNS and Lambda application in Python to trigger the Email anomalies with the help of CloudWatch.

Yahavi-New Delhi, India Nov 2015 – Aug 2017 *Software Developer(Web & Android Application Development)*

- Built the mass mailer campaign with success. Using AWS SES and PHP and sent over 30,000 emails in a day.
- Monitored the mail server using Cloudwatch, SNS, and AWS lambda.
- Optimized the website code for a better load time of about 50% for all the webpages majorly optimizing using Javascript.

Motor & General Sales Pvt. Ltd-Lucknow, India Aug 2014 – Nov 2015 *Software Engineer*

- Maintained company's billing applications, inventories, and hire purchase using C++.
- Built an employee payroll software in wxDev-C++ which is a GUI based and offline desktop-based application for the company.

Education

Master of Science (M.S.), Computer Science
University of Illinois, Chicago
Relevant Coursework: Cloud Computing | Machine Learning | Distributed Systems | AI Safety | Big Data | Deep Learning

Bachelor of Technology (B.Tech), Computer Science and Engineering
Vellore Institute of Technology
Relevant Coursework: OOP | Data Mining | Linear Algebra | Graph Theory | Computer Networks

Projects

- Built a chatbot using Azure AI, LLMs, and KQL to help users optimize queries in Azure Data Explorer. Integrated document embeddings and SQL frameworks to enhance internal data accessibility.

- Developed a GenAI system using AWS Bedrock to forecast stock trends by aggregating data from financial news, websites, and the Federal Reserve.

- Developed an end-to-end RAG system that indexed custom documents using VectorStoreIndex and retrieved contextually relevant data for LLM-based responses. Enabled ability to analyze two companies, from the same sector, using financial earnings release in quarter.

- 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.

- Designed a Investment Agent on Mosaic – Agent Bricks Framework, investment-focused agent using Databricks Agent Framework (MCP). The system orchestrated data retrieval, evaluation, and trade signal generation across real-time market feeds.

- Designed and implemented an end-to-end RAG-based restaurant recommendation system leveraging LlamaIndex, Elasticsearch, and LLMs. Ingested diverse restaurant datasets, embedded metadata for vector search, and deployed a chat-based interface capable of serving personalized suggestions with contextual justifications.

- Engineered a production-ready RAG system using Azure AI Search and LangChain for scalable enterprise knowledge retrieval. Enabled hybrid text–vector search, streaming chat with contextual citations, and Azure-compatible embedding workflows across multilingual documents.