Khalid Rizvi

Enterprise GenAI & Cloud Leader

khalid.rizvi@icloud.com
Vienna, VA - 703-656-6394
linkedin.com/in/khalidrizvi
https://www.khalidrizvi.com

Innovative technology leader with 30+ years of experience architecting and delivering cloud-native, AI-powered solutions for global enterprises and the public sector. Specializing in Generative AI (GenAI) and large-scale modernization, with proven expertise in building RAG (Retrieval-Augmented Generation) systems, orchestrating MCP servers, and deploying LightLLM for high-performance inference at scale. Hands-on experience with Amazon Bedrock embeddings, Knowledge Bases, and agent frameworks to deliver secure, production-ready AI solutions that combine LLMs, embeddings, and advanced prompt engineering. Skilled in ReactJS/AngularJS frontends and high-throughput backends using Go and Java, integrating AI seamlessly into legacy and greenfield systems. Recognized for staying in full control of architecture, security, and compliance while leading AI/ML initiatives that accelerate innovation, reduce costs, and deliver measurable business outcomes.

LEADERSHIP & STRATEGIC VALUE

Enterprise Architecture Leadership:

Led major digital transformations across industries by aligning architecture with business goals to deliver scalable, high-impact solutions.

Innovation and Digital Transformation:

Modernized legacy systems into AI-driven, cloud-native solutions—cut mainframe processing by 99% and delivered 99.9% uptime on critical platforms.

Emerging Technology Integration:

Delivered AI/ML, GenAI, and big data solutions to solve complex challenges—built award-winning modernization tools and AI-powered analytics to drive competitive edge.

Team Leadership and Mentorship:

Built and mentored high-performing teams—fostering innovation, growth, and a culture of excellence through hands-on leadership and collaboration.

Strategic and Crisis Management:

Proven ability to lead high-pressure projects—delivered urgent solutions like COVID-19 systems and federal cloud migrations with speed, quality, and security.

EXPERIENCE

GenAI/AI/ML Solution Engineer *08/2025 - 10/2025*
CTIS – NIH/NCA

GenAI architect delivering advanced AI/ML solutions to modernize oncology research data and enterprise knowledge management for NIH/NCA.

Designed and led the end-to-end architecture of a production-grade RAG platform integrating Amazon Bedrock embeddings, Knowledge Bases, and OpenAI GPT models, with AWS Lex providing conversational AI. Built multi-layered inference and ingest pipelines with FAISS vector search, S3-backed document stores, and LightLLM for optimized, lightweight inference—enabling secure, low-latency semantic search across large oncology datasets. Orchestrated MCP stdio/remote servers for modular agent integration, including Claude Desktop connectivity for hybrid on-premise and cloud workflows. Delivered a fully scalable chatbot framework with session management, prompt lifecycle control, and compliance with NIH data governance. Results included a 70% boost in query accuracy, 55% faster information retrieval, and secure, conversational access to both structured and unstructured datasets.

Associate Partner, Cloud & Enterprise Architect *07/2015 - 06/2025*
DXC Technology

Solution architect overseeing cloud innovation, AI/ML solution delivery, and digital transformation initiatives for major clients in telecom, finance, public sector, and transportation.

Nokia Networks – KPI Analytics & AI Platform (Dec 2024 – Jun 2025): Designed and delivered a next-generation KPI analytics platform for telecom performance monitoring with a Full-Stack ReactJS (latest) SPA integrated into a Single-SPA micro-frontend architecture. Implemented Power BI embedding with client branding and a LangChain + OpenAI GPT chatbot to allow natural language queries over live network data. Backed by high-throughput Go microservices, the system was deployed to Azure Container Apps with Terraform automation (Azure ACR packaging). Achieved a 95% reduction in manual data handling and improved data retrieval speeds by 60%, reducing stakeholder decision-making time from days to minutes. Experience and patterns are directly transferable to AWS micro-frontend architectures.

Network of Giving – Philanthropic Business Cycle Platform (2023 – 2024): Built and supported a ReactJS SPA (Tailwind, Flowbite) with Java/Spring Boot and Go microservices on AWS to handle millions of real-time donation transactions. Delivered complete lifecycle support for the philanthropic business cycle, from donor onboarding to disbursement tracking. Integrated AWS services: Bedrock (Claude v2, Titan), AppSync, DynamoDB, API Gateway, Lambda (Go), ECS/ECR, CloudFront, SES, ACM, and CloudWatch. Features included:

Donor-Cause Matching Engine (Claude v2) to improve donor retention.: Donor Assistant Chatbot for campaign info and giving summaries. Automated Campaign Content (Titan) to reduce NGO overhead. Achieved 99.9% uptime, cut incident resolution time by 70%, and increased development velocity by 50% via DevOps coaching and AWS CI/CD (CodePipeline, CodeBuild, CodeDeploy).

Grants.gov Modernization – Cloud Migration & Security (2022 – 2023): Modernized legacy forms into a ReactJS SPA backed by a Node.js/Express API. Migrated workloads to AWS with CloudFormation/Terraform, S3, Aurora RDS, API Gateway, Lambda (Python), VPC, Route 53, IAM, and CloudWatch/X-Ray. Re-architected the monolith into secure microservices with JWT-based auth and centralized access control. Delivered automated provisioning and deployment pipelines, reducing manual provisioning errors to zero and improving security posture.

Los Angeles Metro – Transit Systems Integration (2020 – 2022): Developed a unified transit monitoring platform, integrating GPS and telematics from 2,000+ buses and trains into a central fleet dashboard. Delivered ReactJS frontends and Java integration services with Dell Boomi. Achieved 15% fuel cost savings (~$2M/year) and reduced parts delays by 30%. Implemented predictive analytics for maintenance, cutting service disruptions.

NY Metropolitan Transportation Authority – Enterprise Architecture & Systems Modernization (2018 – 2020) Delivered multiple modernization initiatives:

- Asset Management Integration (OAGIS ↔ MIMOSA Transformation): Built a ReactJS SPA for schema validation and transformation, backed by Spring Boot microservices. Unified Infor EAM and Bentley AssetWise via microservices on Red Hat OpenShift, translating between MIMOSA and OAGIS. Resulted in a 40% reduction in asset downtime.
- Electric Bus Predictive Maintenance & Energy Optimization: Used AWS SageMaker to train predictive models on telemetry and charging data for 500+ buses, reducing battery failures by 20% and improving maintenance scheduling accuracy by 35%.
- COVID-19 Sanitation Compliance: Created a high-throughput sanitation tracking system using Java, Apache Camel, and a ReactJS UI for real-time compliance status—ensuring 100% regulatory compliance.

Mainframe Batch Modernization – Daito Corp (2017 – 2018): Migrated AS/400 COBOL batch processing to Java Spring Boot, cutting runtime from 23 hours to 13 minutes (99% improvement).

Mastercard – Global Loyalty Platform Integration (2016 – 2017): Rebuilt loyalty integration platform with Spring Boot microservices, improving throughput by 60% and enabling rapid partner onboarding.

Senior Solutions Architect *11/2006 - 07/2015*
RIZ Consulting

- Led high-impact federal IT modernization projects, designing and delivering advanced solutions for agencies like EPA, VA, USDA, and Grants.gov.
- Engineered automated rules engines and modern web applications that increased operational efficiency by up to 65% and improved data accuracy, compliance, and user satisfaction for millions of users.

Senior Architect *07/1997 - 11/2006*
CSC Consulting

- Led multi-million dollar software modernization and integration projects for Fortune 500 clients, accelerating delivery by developing reusable frameworks and automation tools.
- Architected secure, high-availability platforms in finance, manufacturing, education, and healthcare, reducing development timelines by 70% and ensuring zero production downtime.

EDUCATION

Master of Science in Computer Science - California State University, Sacramento *01/1995 - 07/1997*

Master of Science in Mechanical Engineering - NED University of Engineering & Technology *01/1986 - 01/1988*

Bachelor of Science in Mechanical Engineering - NED University of Engineering & Technology *01/1982 - 01/1986*

MODERN TECHNICAL SKILLS

Programming & Scripting:

Python, Go, Java, Javascript, Typescript, Bash, SQL

Web & SPA Frontend

Frameworks & Architectures: React (SPA), Next.js (SSR/SSG/ISR), Angular and AngularJS (legacy), Jamstack with Hugo, Go HTML Templates for server-side rendering.

React Ecosystem: Hooks, Context API, React Router, Redux Toolkit, Zustand, TanStack Query (server-state caching), React Hook Form, code splitting, lazy loading, Vite/Webpack builds.

UI & Styling: Tailwind CSS (utility-first), Flowbite component library, responsive/mobile-first CSS, semantic HTML5, accessibility (WCAG), cross-browser compatibility.

Patterns & Performance: SPA state normalization, client vs. server rendering trade-offs, bundle size control, memoization and virtualization, API integration (REST/GraphQL), environment-based configuration and CI/CD for modern frontends.

Cloud Platforms:

AWS (EC2, S3, RDS, Lambda, API Gateway, SageMaker, Bedrock, Amazon Q), Azure (Azure ML, Container Apps), Google Cloud

Machine Learning & AI:

Skilled in building and tuning machine learning models like regression, clustering, neural networks, and NLP using transformers such as BERT and RoBERTa.

Generative AI Frameworks & Tools:

Experienced with tools like NeMo, Whisper, LangChain, LlamaIndex, Haystack, and Streamlit to build secure and interactive generative AI apps. Skilled in deploying LightLLM, MCP servers, Amazon Bedrock (embeddings, Knowledge Bases, Guardrails), Hugging Face Transformers, and OpenAI APIs, integrating them into production-grade RAG pipelines and enterprise chatbot frameworks.

Frameworks & Libraries:

TensorFlow, PyTorch, Keras, scikit-learn, Pandas, NumPy

Big Data & Analytics:

Apache Spark, Apache Kafka – for large-scale data processing; Tableau, Power BI – for data visualization and dashboards

Databases:

SQL (MySQL, PostgreSQL, Oracle) and NoSQL (MongoDB, DynamoDB, Elasticsearch, Redis)

DevOps & MLOps:

Docker, Kubernetes, Jenkins, CI/CD pipelines; Infrastructure as Code (AWS CloudFormation, Terraform, AWS SAM); Configuration management (Ansible). Skilled in end-to-end MLOps lifecycle management from model experimentation and versioning to production deployment and monitoring.

WebAssembly & Edge

WebAssembly (WASM): for compute-heavy browser features; integration with JavaScript and Go (TinyGo), loading .wasm modules via Vite/Webpack.

Tooling & Runtimes: AssemblyScript, Emscripten, WASI basics; edge platforms such as Cloudflare Workers, Vercel Edge Functions, and Deno Deploy.

Security & Compliance:

OAuth 2.0, mTLS encryption, JWT authentication

CERTIFICATIONS

Talk to Your Documents with LangChain and Python *- 08/2025*

MCP Servers Made Easy with Python and OpenAI Agents *- 07/2025*

AWS Certified Cloud Practitioner *- 08/2024*