

# NIKHIL KUMAR TAMMI

[nikhiltammi.tech@gmail.com](mailto:nikhiltammi.tech@gmail.com) • (425) 419-1237 • [linkedin.com/in/nikhiltammi/](https://linkedin.com/in/nikhiltammi/) • [github.com/nikhiltammi/](https://github.com/nikhiltammi/)

## SUMMARY

AI Software Engineer & Cloud Architect with 5+ years of experience designing scalable Generative AI platforms and optimizing enterprise cloud infrastructure across AWS, Azure and GCP. Proven expertise in building production-grade multi-agent workflows using AWS Bedrock and MCP, migrating 7M+ compliance-critical files, and implementing secure, non-hallucinating RAG solutions. Strong background in architecting cost-efficient CI/CD pipelines and operationalizing advanced AI systems into secure, compliant, and high-performance production platforms.

## SKILLS

**Programming & Scripting:** Python, Java, SQL, Bash

**Generative AI & Agent Systems:** AWS Bedrock, AgentCore, MCP, RAG, GPT, BERT, Hugging Face, LangChain, CrewAI, Strands, Multi-Agent Systems

**DevOps & Platform Engineering:** AWS CDK, Terraform, Kubernetes, Docker, Carpenter, Fargate, Jenkins, GitHub Actions, CodePipeline

**Cloud & Infrastructure:** AWS (EC2, EKS, ELB, RDS, S3, Lambda, SageMaker, CloudFormation), GCP, Azure

**Monitoring & Logging:** AppDynamics, Cloudwatch, Datadog, Prometheus, Grafana

**Source & Artifact Management:** GitHub Enterprise, Bitbucket, GitLab, Artifactory, Maven, GitLFS, Git (Forking, Gitflow)

**Data & Streaming:** Kafka, Apache Flink, Spark, Snowflake, Data Lake Formation, SQL Server, Oracle DB, DynamoDB

**Machine Learning Platforms & APIs:** PyTorch, TensorFlow, Keras, AWS Textract, MLflow, FastAPI, REST API Development

## PROFESSIONAL EXPERIENCE

**PEOPLE TECH GROUP INC – Sr AI Software Engineer & Cloud Consultant**

Nov 2023 – Present

**CLIENT: BOEING**

Sep 2025 – Present

- Designed and implemented an AI-powered export document delivery platform, migrating 150+ file formats to AWS while improving regulatory compliance efficiency by 30%.
- Led the large-scale migration of 7+ million files from on-prem systems to Amazon S3, preserving 100% ACL fidelity through precise permission and entitlement mapping.
- Engineered an automated post-migration file identification and classification pipeline, enabling accurate categorization and faster downstream compliance workflows.
- Built a persona-based application layer (admin, reviewer, end user) that mirrored on-prem functionality, reducing file access issues and improving operational efficiency by 30%.
- Implemented secure Smart Search using Amazon Kendra integrated with Amazon Bedrock, enabling authorized users to retrieve summaries and insights from large documents while enforcing strict access controls.
- Designed the platform with security-by-design principles, ensuring role-based access, auditability, and adherence to export control standards.
- Collaborated with cross-functional compliance and platform teams to develop the solution, supporting enterprise-scale usage with high availability and governance.
- Implemented end-to-end audit logging and traceability for file access and compliance actions, enabling faster investigations, simplified audits, and improved regulatory reporting readiness across the platform.

**CLIENT: QUINTSTREET**

Jan 2025 – Sep 2025

- Architected and built a production-grade GenAI automation system using Strands Agents SDK, leveraging multi-agent orchestration (GitHub Agent, Transformation Agent, Execution Agent) to generate and deploy denormalized views in Snowflake, fully integrated with GitHub and Snowflake MCPs.
- Developed an enterprise GenAI financial chatbot, generating optimized SQL scripts for natural-language prompts, enhancing data retrieval by 30% and supporting 10+ visualization formats, thereby improving decision-making efficiency across finance teams.
- Implemented an initial MVP using Amazon Bedrock and LangGraph to orchestrate conversational agents for natural-language-to-SQL analytics, accelerating agent workflows by 40% and enabling rapid validation with 100+ internal users.
- Migrated the MVP to a scalable, production-grade multi-agent architecture using Bedrock AgentCore, improving system scalability by 50% and ensuring reliable performance under high query volumes.
- Optimized Bedrock AgentCore Runtime and Memory with Strands and Redshift MCP, achieving 25% faster data processing and secure, governed agent-to-database communication.
- Integrated Active Directory-based identity and role mapping with dynamic runtime prompt injection, enforcing fine-grained data access controls for 500+ users.
- Delivered a non-hallucinating, guardrail-enforced conversational engine with streaming responses and <15s average latency for complex analytical SQL queries.

**CLIENT: SEIU 775 BENEFITS GROUP**

Nov 2023 – Jan 2025

- Led multiple enterprise cloud, DevOps, and GenAI initiatives spanning CI/CD optimization, database automation, data modernization, and legacy system migration, delivering measurable gains in deployment speed, cost efficiency, and platform reliability.
- Optimized large-scale CI/CD pipelines by refactoring 36 standalone CloudFormation stacks into nested stacks, reducing template size from ~950 KB to ~300 KB and cutting deployment time from ~3 hours to ~1.5 hours.
- Redesigned artifact handling in CodePipeline by uploading build assets to S3 and referencing relative paths, eliminating CloudFormation template bloat and improving pipeline stability.
- Implemented database CI/CD automation using Bytebase, decoupling schema changes from application code and reducing production schema deployment time from 3–4 hours to under 1 minute.
- Built a GenAI-driven Snowflake automation pipeline that converts stored procedures into denormalized analytical views using GitHub MCP and Snowflake MCP, executing and validating transformations automatically.
- Designed an agent-orchestrated modernization platform to convert Oracle PL/SQL into AWS Glue Java jobs, enabling automated extraction, conversion, validation, deployment, and execution through a single workflow.
- Applied GenAI-assisted validation and automation to data transformation and migration workflows, reducing manual verification effort and ensuring consistent, safe execution across CI/CD-driven modernization pipelines.

## SS&C TECHNOLOGIES

July 2021 – July 2022

### Software Engineer

- Led migration of enterprise GitLab CI/CD infrastructure from OCI OKE to AWS EKS, modernizing pipelines supporting 300+ servers across four environments.
- Replaced always-on runners with Kubernetes-based GitLab executors, leveraging Karpenter, Fargate, and Graviton EC2 for dynamic, right-sized compute provisioning.
- Reduced CI/CD compute costs by ~60% through dynamic, on-demand scaling by enabling on-demand scaling (light jobs on small nodes, heavy builds on larger instances, burst workloads on Fargate).
- Introduced a hybrid GitLab runner architecture with an EC2-based runner host to support IAM role integration and specialized workloads (e.g., Docker-in-Docker).
- Implemented drift detection using CDK plan/state validation, preventing configuration mismatches across environments and improving deployment reliability.
- Embedded GDPR and NIST compliance checks into CI/CD pipelines using AWS Conformance Packs, cfn-nag, and checkov, optimized to enforce security without slowing delivery.

## M-DIGITAL TECH

Dec 2019 – May 2021

### Software Engineer

- Designed and deployed modular AWS infrastructure using Terraform and Ansible, provisioning 10+ VPC components (VPCs, subnets, EC2, IAM roles, security groups) to support scalable ML environments across multiple stages (dev/test/prod).
- Architected and deployed cloud-native ML pipelines using AWS ECS, Lambda, and S3, enabling scalable training and inference workflows with containerized Python models.
- Engineered scalable ML pipelines and CI/CD workflows using Python, Docker, ECS, and Jenkins—ensuring reproducibility and fault-tolerant deployments in dynamic environments.
- Built a real-time data ingestion framework using Amazon Kinesis, S3, and Redshift, supporting high-throughput log processing and downstream analytics. Managed microservices deployments using Kubernetes Helm charts, enabling zero-downtime releases, versioned rollbacks, and consistent configuration across 10+ services.
- Integrated SonarQube SAST into CI/CD pipelines, enforcing automated code quality checks and vulnerability scanning on 100% of code commits.
- Set up cross-region replication for S3 and DynamoDB to ensure disaster recovery readiness and regional fault tolerance. Managed cloud lifecycle policies for EBS and S3 backups, integrated with AWS Backup and encrypted using KMS-managed keys for data protection.

## EDUCATION

### UNIVERSITY OF CINCINNATI

Master of Engineering in Artificial Intelligence

Cincinnati, OH

Apr 2024

### MALLA REDDY ENGINEERING COLLEGE

Bachelor of Technology in Computer Science and Engineering

Hyderabad, Telangana

July 2021

## PROJECTS

### Strands GenAI Automation for Snowflake Denormalized View Generation

- Architected a GenAI-driven automation system using Strands Agents SDK for Snowflake denormalized view generation.
- Designed and implemented multi-agent orchestration with GitHub, Transformation, Execution, and Snowflake Agents.
- Integrated GitHub MCP and Snowflake MCP to enable secure, modular, and dynamic agent execution.
- Automated complex SQL generation and execution workflows, reducing manual effort and improving data reliability.

### Test Case Automation using AWS Bedrock

- Designed a GenAI-driven automation framework leveraging AWS Bedrock and multi-agent collaboration to generate and execute test cases.
- Enhanced testing efficiency by automating validation processes, reducing manual effort, and improving software reliability.

### Chat with CI/CD Pipelines

- Developed an AWS Bedrock and GenAI-enhanced chat system to streamline CI/CD processes, boosting deployment efficiency.
- Refined CI/CD event handling dynamic insights and tailored recommendations, improving overall pipeline management quality.

### GenAI Model and RAG Evaluator

- Implemented a GenAI-powered system on AWS Bedrock to evaluate RAG models, enhancing model selection and system scalability.
- Optimized RAG model evaluation by synthesizing file summaries and output analysis, securing optimal performance with AWS integration.

## CERTIFICATIONS (Active)

- GCP Cloud Architect – Professional
- AWS Certified Solution Architect – Associate
- AWS Certified Developer – Associate
- Google Cloud Associate Cloud Engineer
- Microsoft Certified: Azure Fundamentals