



Muhammad Husnain Ali

AI Solution Architect

AI Consultant | Azure, AWS, GCP

ABOUT ME

AI Solution Architect and AI Consultant with 5 years of experience designing cloud-native, enterprise-grade AI and banking platforms across Microsoft Azure, AWS, and Google Cloud (GCP). Architects event-driven, microservices, and modern monolith solutions, and turns business goals into governed, production AI for UAE banking, energy, and government clients, including First Abu Dhabi Bank, Dubai Islamic Bank, RTA, TAQA, and Fertiglobe.

EDUCATION

Bachelor of Science (B.Sc.), Computer Science

Superior University, Lahore

CGPA: 3.6/4.0

CERTIFICATIONS

- Microsoft Copilot Studio & Azure AI
- LangChain & LangGraph: Agentic AI
- Azure AI: Cloud-Native Architecture
- Prompt Engineering & LLM Optimization
- AWS Certified Machine Learning Engineer Associate (MLA-C01), Expected Q3 2026

TECHNICAL SKILLS

Solution Architecture

Event-Driven Architecture (EDA), Microservices Architecture, Modern Monolith / Modular Monolith, Domain-Driven Design (DDD), CQRS, Event Sourcing, Saga & Outbox Patterns, API-First Design, Distributed Systems, System Design, Enterprise Solution Design, Reference Architectures

Microsoft Azure

Azure AI Foundry, Azure OpenAI, Azure Machine Learning, Copilot Studio, Azure Kubernetes Service (AKS), Container Apps, Azure Functions, Event Hubs, Service Bus, API Management, Cosmos DB, Azure AI Search, VNet, Private Endpoints, Managed Identity, Key Vault, Microsoft Entra ID

Amazon Web Services (AWS)

Amazon Bedrock, SageMaker, Amazon EKS, ECS, AWS Lambda, API Gateway, Amazon MSK (Kafka), EventBridge, SQS / SNS, Amazon S3, RDS, DynamoDB, IAM, VPC

Google Cloud (GCP)

Vertex AI, Gemini, Google Kubernetes Engine (GKE), Cloud Run, Cloud Functions, Pub/Sub, BigQuery, Cloud Storage, Cloud SQL, IAM

+92 300-0916-426

linkedin.com/in/muhammadhusnainali873

muhammad.husnain.ali.738@gmail.com

github.com/MuhammadHusnainAli

USD 50M+ Revenue Impact

73% RAG Latency Reduction (15s to 4s)

1000+ Hours Saved Annually

3 Clouds: Azure, AWS, GCP

5 Years Architecture & AI Delivery

25+ Production AI Solutions

95-98% Eval Accuracy (Enterprise)

65%+ Cross-Department Productivity

WORK EXPERIENCE

AI Solution Architect & AI Consultant (AI / ML Engineer)

DigiFloat
Lahore, Pakistan

Nov 2024 – Present

- Lead solution architect and client-facing AI consultant for UAE/GCC banking, energy, and government clients: run discovery and requirements workshops, define target-state architectures, record Architecture Decision Records (ADRs), and steer delivery from PoC to production.
- Architected **banking cloud solutions** for First Abu Dhabi Bank and Dubai Islamic Bank that combine a **modern (modular) monolith** for the transactional core, **microservices** for independently scaling workloads, and **event-driven architecture** on **Kafka / Azure Event Hubs**, secured with private networking, Entra ID, and audit trails.
- Design multi-cloud reference architectures on **Microsoft Azure** (AKS, AI Foundry, API Management), **AWS** (Bedrock, EKS, Lambda, MSK), and **Google Cloud** (Vertex AI, GKE, Cloud Run, Pub/Sub), choosing services on cost, latency, data residency, and compliance.
- Delivered 25+ agentic AI systems with LangChain, LangGraph, Azure AI Services, and Copilot Studio; pioneered production **MCP** (Model Context Protocol) and **A2A** (Agent-to-Agent).
- Set up MLOps / LLMOps on Azure DevOps (CI/CD, model versioning) with Helm releases to **AKS**, Container Apps, and Functions, and Langfuse for tracing and evaluation.

Key Projects:

- Data2Decisions (Enterprise Data Intelligence):** Multi-agent text-to-insight platform built around a central orchestration agent that coordinates 10+ specialised sub-agents, each with its own scoped skills. Unifies unstructured sources (SharePoint, S3, Google Drive) with 25+ structured connectors (MySQL, Oracle, MongoDB) over a unified knowledge-graph layer that powers cross-source reasoning. Data governance and access controls scope every sub-agent to its permitted domain. **USD 50M+ revenue impact**; 1000+ engineering hours saved annually.
- NovaFloat (Enterprise AI Productivity Platform):** AI platform unifying **1500+ prebuilt connectors**, custom Python **MCP servers** attached from the UI, and an n8n-style workflow builder, with a **harness-engineered agent** reasoning across connectors, knowledge bases, and live web search. **FastAPI** backend with **Microsoft Entra ID** OAuth, PostgreSQL, **Azure AI Search**, and Redis; lifted cross-department productivity **65%+**.
- Cheque Clearance Automation (First Abu Dhabi Bank, Banking Cloud):** Event-driven microservices platform on **Kafka / Azure Event Hubs** where partitioned topics stream cheques through microservice consumers: image preprocessing, **YOLO** detection, OCR extraction, signature matching, then automated clearance. Tuned topic partitioning and consumer groups to remove stalling under load; deployed on **AKS** with Helm, private networking, and autoscaling for secure, bank-grade throughput, with idempotent consumers, retries, and dead-letter topics so no cheque is lost or processed twice.
- Plutus (Dubai Islamic Bank):** Conversational banking agent for balances, transfers, and beneficiaries. Wrapped core banking REST APIs into a custom **MCP server** exposed to CrewAI agents, applying a layered governance model of guardrails, both AI-based and hard-coded, enforced at tool and agent level (e.g. balance replies constrained to the single numeric value the tool returns), keeping every response within regulated, auditable banking boundaries.
- Fabric Extraction Pipeline (Fertiglobe UAE):** End-to-end Microsoft Fabric pipeline ingesting policy PDFs from SharePoint and extracting 40+ structured fields into Delta tables. Azure AI Foundry agents with JSON-schema response formats drive a multi-step per-PDF extraction plus a dedicated evaluation agent for sensitive financial figures, reaching **95% extraction accuracy and 98.75% evaluation accuracy**.
- CFO Digest (UAE Client):** Automated multi-stage AI + data pipeline, end-to-end from scrape to email. Serper API pulls date-scoped market/financial data into Fabric warehouse tables; downstream agents load and threshold-filter records, draft stakeholder narratives, then hand off to an evaluation agent. Runs weekly to generate a Power BI PDF report emailed per stakeholder group, orchestrated via an Azure AI Foundry agent.

Banking & FinTech Cloud
Core Banking API Integration, Payments & Cheque Clearing, Event Streaming, Apache Kafka, Zero-Trust & Private Networking, Regulatory Compliance, Audit Trails, High Availability, Disaster Recovery
Consulting & Delivery
Stakeholder Workshops, Requirements Discovery, AI Strategy & Roadmaps, Architecture Decision Records (ADR), PoC to Production, Cost Optimisation (FinOps), Well-Architected Reviews, Pre-Sales & Solutioning, Technical Leadership
Generative AI
Generative AI, Large Language Models (LLM), Agentic AI, Multi-Agent Orchestration, Sub-Agents & Skills, Harness Engineering, Context Engineering, Prompt Engineering, Retrieval-Augmented Generation (RAG), Fine-Tuning, LoRA/QLoRA, Embeddings
LLM Stack & Vector Search
LangChain, LangGraph, CrewAI, GPT-4o, Claude, Gemini, LLaMA, Hugging Face, Pinecone, FAISS, Chroma, Weaviate, Neo4j, Knowledge Graphs, Hybrid Search
DevOps, MLOps & IaC
Kubernetes, Helm, Docker, Terraform, Azure DevOps, GitHub Actions, CI/CD, MLOps, LLMOps, MLflow, Langfuse (Tracing, Monitoring), Evaluation Pipelines
Data Engineering
Microsoft Fabric, ETL / ELT Pipelines, Delta Tables, Azure Data Factory, Lakehouse, Data Warehouse, SQL, Text-to-SQL, Power BI
Protocols & Integration
MCP (Model Context Protocol), A2A (Agent-to-Agent), REST APIs, GraphQL, Webhooks, OAuth 2.0, SharePoint, MS Teams, HubSpot CRM
Security & AI Governance
Guardrails, Access Control (RBAC), Responsible AI, Data Governance, Evaluation & Grounding, Auditability, Secrets Management
Backend, ML & Vision
Python, FastAPI, PostgreSQL, Redis, Git, PyTorch, TensorFlow, Machine Learning, Deep Learning, Natural Language Processing (NLP), YOLO, OCR, Document Extraction
INDUSTRIES
Banking / FinTech Cloud (FAB, DIB)
Government (RTA UAE, Data2Decisions)
Energy (Fertiglobe UAE: DOA, VAT)
Legal Tech (LawGPT)
Healthcare (MedConverse, HealChat)
HR Tech (Screening, Email Automation)
Enterprise Productivity (NovaFloat)

- **CFO Board Agent (Fertiglobe UAE):** Board-level Copilot Studio agent that distils source content into concise 3-line briefs from clear structured instructions, built on secure document grounding with a dedicated evaluation loop, reaching **87% evaluation accuracy**.
- **RTA Mahboob (UAE Govt, 1000+ users):** Multilingual RAG (EN/AR) that routes each query by language into a dedicated English or Arabic pipeline for low latency, alongside text-to-SQL over 500+ tables via FastAPI. Cut response latency from 15s to 4s (**73%**).
- **UAE VAT Agent (Fertiglobe UAE):** FTA-reference, amendment-aware tax-guidance agent grounded on curated FTA source documents for verifiable, citable answers. Deployed to Microsoft Teams. **98% accuracy** across 100+ eval cases.
- **DOA (Fertiglobe UAE):** Policy Q&A via Copilot Studio. Source PDFs encoded values as colour coding rather than literal numbers, so a preprocessing pipeline converts colour-coded documents into clean Markdown before the agent sees them, fixing the data problem ahead of the agent problem. SharePoint + Entra ID integrated; **95% accuracy** on 200+ eval cases.
- **LawGPT (Legal AI, Pakistan):** Platform for lawyers and judges spanning legal chat, document drafting, and document analysis, where a centralised orchestrator agent delegates to feature-specific sub-agents on demand. 300K+ legal cases indexed.
- **Tanzim (Driver Support Automation):** AI support platform covering 5 to 6 driver workflows (lost-and-found, fare disputes, and more). Each inbound email triggers a LangGraph workflow in FastAPI that resolves autonomously or escalates by auto-creating a **HubSpot CRM** ticket, under layered guardrails. Secured behind an **Azure API Management** gateway, with Container Apps and the **MCP server** isolated inside a private VNet.

Data Scientist / ML Engineer	
<i>Swift Solver</i> <i>Remote</i>	Jan 2024 – Jul 2024
• Built end-to-end ML pipelines across 50K+ records; fine-tuned SLMs with LoRA/QLoRA for a 25% accuracy gain. • Built NLP classification and sentiment pipelines using Transformers and BERT; set up model versioning, experiment tracking, and automated retraining.	
Python Developer / AI Engineer	
<i>Pure Logics</i> <i>Lahore, Pakistan</i>	Jun 2023 – Dec 2023
• Fine-tuned GPT-3.5-turbo via OpenAI API with a 40% accuracy gain; built preprocessing pipelines across 100K+ documents. • Shipped FastAPI and Streamlit apps in production with auth, rate limits, and logging; applied few-shot and chain-of-thought prompting.	

SOLUTION ARCHITECTURE CAPABILITIES
• Event-Driven Architecture: Kafka, Azure Event Hubs, Service Bus, AWS MSK / EventBridge, and GCP Pub/Sub; topic and partition design, consumer groups, event sourcing, outbox and saga patterns for reliable, auditable banking transaction flows. • Microservices Architecture: domain-driven service boundaries, API gateways (Azure API Management, AWS API Gateway), containerised services on AKS, EKS, and GKE with Helm, autoscaling, and zero-trust private networking. • Modern (Modular) Monolith: well-bounded modules in a single deployable for transactional banking cores and fast-moving products, keeping strong consistency and low operating cost, with a clear path to extract microservices as scale demands. • Multi-Cloud AI Platforms: Azure AI Foundry / Azure OpenAI, Amazon Bedrock / SageMaker, and Google Vertex AI / Gemini, with RAG, agentic workflows, MCP servers, evaluation, and guardrails. • Consulting & Governance: AI strategy and roadmaps, well-architected and cost reviews, security and compliance by design (RBAC, Entra ID / IAM, Key Vault), and stakeholder-ready documentation.
KEY ACHIEVEMENTS

- Architected banking cloud solutions for **First Abu Dhabi Bank** and **Dubai Islamic Bank**, combining event-driven, microservices, and modern-monolith patterns on **Kafka / Event Hubs** and **AKS**.
- Drove **USD 50M+ revenue impact** and saved 1000+ engineering hours annually through the Data2Decisions multi-agent platform.
- Achieved **95% extraction** and **98.75% evaluation accuracy** on the Fertiglobe Microsoft Fabric pipeline, and cut RTA Mahboob RAG latency by **73% (15s to 4s)** for 1000+ users.