

# Muntazir Fadhel

ML Systems Architect · Founder, HADI Technology · Toronto · Najaf · Dubai

SITE [mfadhel.com](https://mfadhel.com)

GIT [github.com/hadi-technology](https://github.com/hadi-technology)

IN [linkedin.com/in/muntazirfadhel](https://linkedin.com/in/muntazirfadhel)

## SUMMARY

**Senior ML systems architect, 8+ years across enterprise banks, AI research labs, and AI-native startups.** Architect end-to-end ML and GenAI systems, from data ingestion through model serving, MLOps, and monitoring, on AWS, GCP, and Kubernetes. A decade of ML infrastructure shapes how I approach LLM systems today. Built and operate **striff.io**, a live LLM + GNN production system. Led hybrid Kubernetes migration at **RBC Royal Bank** under bank change management. Built ML research infrastructure at **BorealisAI** on NVIDIA DGX clusters.

## Core Capabilities

*Production · public reference implementations*

### LLM & GenAI Systems · RAG · Embeddings

- › Production **vLLM serving on Kubernetes via KubeAI** with scale-from-zero, KV-cache routing, and OpenAI-compatible APIs (self-hosted LLM stack write-up).
- › Embedding-based retrieval, RAG with traceable outputs, and hybrid integration of self-hosted models (DeepSeek, Llama 3) with frontier APIs (GPT, Claude, Bedrock).

[hadi-technology/vllm-mlops](https://github.com/hadi-technology/vllm-mlops)

### Graph RAG · Neurosymbolic · NLP · Model Optimization

- › Graph RAG with **ShEx-constrained schema validation**, phased validator-LLM feedback, and cross-document merge for accuracy and traceability where pure vector RAG fails (Fylo neurosymbolic pipeline write-up).
- › Distilled R-GCN scoring on Triton Inference Server with ONNX Runtime; Transformer-based NLP/NER pipelines for entity and sentiment extraction at production scale.

[hadi-technology/striff-gnn](https://github.com/hadi-technology/striff-gnn)

### End-to-End MLOps · CI/CD/CT · Monitoring

- › End-to-end pipelines with **MLflow** tracking, gated promotion, **Argo CD** GitOps, and blue/green model rollouts via Argo Rollouts (production MLOps reference architecture).
- › Drift detection, model versioning, automated retraining hooks, and Vault-backed governance. Promotion is a merge, not a redeploy.

[hadi-technology/mlops-blueprint](https://github.com/hadi-technology/mlops-blueprint)

### Cloud-Native Platforms · K8s · Terraform · GPU

- › Multi-cloud ML platforms across **AWS, GCP, and on-prem Kubernetes**; hybrid container infrastructure under bank change management; Terraform-driven IaC.
- › GPU orchestration with autoscaling and continuous batching; observability with Prometheus, Grafana, EFK; compliance-aligned for OSFI, change advisory boards, audit-grade deployment.

**RBC Royal Bank** · 2021-2022 · OSFI-aligned · **BorealisAI** · NVIDIA DGX

## PRODUCTION-GRADE

ML systems serving live workloads, not POCs. Async pipelines, autoscaled inference, blue/green rollouts under sustained load.

## ENTERPRISE CONTEXT

Built under bank change management at **RBC**; ML research compute at scale at **BorealisAI** on NVIDIA DGX clusters with SLURM orchestration.

## TECHNICAL LEADERSHIP

Founder and lead engineer at HADI Technology. Lead architect across 7+ active client engagements. Peer-reviewed research at IEEE-affiliated venue.

## Reference Project

*Live · public · end-to-end*

### striff.io → [striff.io](https://striff.io)

[end-to-end](#) [live](#) [open-core](#)

A live AI architectural-review system for GitHub pull requests, combining classical ML, deep learning, and LLM-based reasoning in a single end-to-end production pipeline. Webhooks land in a Kafka-staged async pipeline with three independent worker tiers (graph construction, GNN scoring on **Triton Inference Server**, and LLM annotation), each scaling on its own characteristics.

A three-tier degradation hierarchy means every failure mode produces a weaker but still useful output rather than an error: full pipeline → symbolic-only → retroactive-symbolic. **Distilled R-GCN** with a 404-dim feature vector (knowledge

## STACK

- › Java · Python
- › ONNX Runtime · Triton
- › Kafka · MongoDB
- › Argo CD · Argo Rollouts
- › Prometheus · Grafana
- › Kubernetes · Terraform

## PUBLIC ARTIFACTS

[hadi-technology/striff-lib](https://github.com/hadi-technology/striff-lib)

distillation for production latency), blue/green model rollouts via Argo Rollouts, contract tests between feature builder and ONNX model.

The underlying approach builds on research published at **ICCC 2021 in cooperation with IEEE Computer Society**. [striff-lib](#) is open source and on Maven Central as `io.github.hadi-technology:striff-lib`.

- [hadi-technology/striff-gnn](#)
- [hadi-technology/clarpse](#)
- [mfadhel.com/striff-io-ml-infrastructure](#)

## Track Record

Verifiable engagements

2021 – 2022 · Contract

### RBC Royal Bank *Cloud Infra Eng.*

Migration of a critical Apigee Edge platform to hybrid AKS + Anthos under bank change management. GitOps with Flux/Kustomize, automation framework for cluster ops and credential rotation, chaos testing of runtime services.

*Apigee · AKS · Anthos · Flux · Ansible · OSFI*

2019 – 2020 · Full-time

### BorealisAI *ML Infra Dev.*

RBC's AI research lab. SLURM on NVIDIA DGX clusters for research compute. PureStorage S3 adoption with 100%+ ingestion throughput gain. OpenShift CI/CD for ML deployments with Jenkins, Vault, Artifactory.

*SLURM · NVIDIA DGX · OpenShift · Jenkins · Vault · Artifactory*

2018 – 2019 · Full-time

### Buzz Indexes *MLOps Eng.*

NLP extraction pipelines turning unstructured financial text into structured entity and sentiment signals powering live equity indexes. Migrated to event-driven serverless on AWS, cutting pipeline runtime by 50%.

*AWS · ECS · Lambda · NLP · entity extraction*

## Recent Client Engagements

via HADI Technology · 2023+

### SkanAI ↗

Built ML inference infrastructure for financial services intelligence; reduced model deployment time to minutes.

### Joinable ↗

Designed and deployed self-hosted LLM serving on Kubernetes; eliminated reliance on external inference APIs.

### FyloAI ↗

Built scientific graph RAG and neurosymbolic extraction pipeline with ShEx-constrained LLM loops.

### ZahraTrust ↗

Architected AI agent infrastructure with traceable inference and controlled tool-use boundaries.

### BrandWell ↗

Operationalized content ML systems with reproducible MLflow workflows and gated model promotion.

### Moxby

Stood up ML platform foundation: cross-cloud pipelines, artifact versioning, observability stack.

### Terradotta ↗

Designed ML pipeline operations with drift monitoring and structured human-in-the-loop validation.

## Engagement Modes

Three ways to start

### 01 · Assess

#### Platform Review

Current state, target architecture, GPU economics, gaps. Written report and a working session.

*3–4 weeks · fixed-price*

### 02 · Build

#### Implementation

Self-hosted LLM serving, ML platforms, async pipelines, observability, governance.

*8–16 weeks · outcome-priced*

### 03 · Operate

#### Managed Platform

Ongoing operations, model lifecycle, GPU cost optimization, on-call coverage.

*monthly retainer*

Contact [mfadhel.com/contact](#) or via this Upwork conversation

brief.v.2026.05