

Hussein Chedi

Dallas, TX | hussanichedi@outlook.com | +1 (214) 519-9068

SUMMARY

AI/ML Solutions Architect with 10 years of experience designing machine learning systems on cloud infrastructure. Skilled in Python, PyTorch, Kubernetes and multi-GPU distributed training, cutting inference latency by **40%** and GPU costs by **30%**. Thrives in remote-first workflows, collaborating across distributed engineering and product teams.

TECHNICAL SKILLS

Programming Languages: Python | Go | C++ | Java

ML Frameworks: PyTorch | TensorFlow | JAX | Hugging Face

Distributed Training & Inference: DeepSpeed | NCCL | vLLM | Triton Inference Server

GPU Cloud Infrastructure: NVIDIA Hopper | CUDA | Slurm | AWS

Containers & DevOps: Docker | Kubernetes | Terraform | Git

Remote Collaboration: GitHub | Jira | Slack | Async Communication

WORK EXPERIENCE

AI/ML Solutions Architect

IBM | Dallas, TX (Remote) | Sep 2022 - Aug 2026

- Architected distributed training pipelines using **PyTorch** FSDP and Kubernetes on multi-GPU clusters, cutting model training time by **35%** across 12 enterprise workloads.
- Optimized LLM inference serving with vLLM and Triton on NVIDIA GPUs, lowering p95 latency by **42%** and reducing annual GPU spend by \$180K for customers.
- Designed reference architectures and documentation for Kubernetes-based MLOps platforms, helping enterprise customers adopt and scale AI workloads on hybrid cloud.
- Built Terraform modules and Helm charts provisioning GPU node pools and Slurm clusters, cutting environment setup from 3 days to 4 hours for customer teams.
- Delivered customer workshops, demos and tutorials on distributed training and Hugging Face fine-tuning, translating GPU infrastructure concepts for engineering and product teams.
- Troubleshoot NCCL communication bottlenecks and CUDA memory issues in multi-node training jobs, improving GPU utilization from 58% to **87%** for large language model workloads.

Machine Learning Engineer

Devsu | Orlando, FL (Remote) | Jan 2020 - Aug 2022

- Developed PyTorch and TensorFlow models for recommendation and NLP use cases, improving prediction accuracy by **18%** for 6 fintech and retail clients.
- Containerized ML training and inference services with **Docker** and **Kubernetes**, reducing deployment time by **60%** and supporting 2M daily predictions.
- Built MLOps pipelines with MLflow, GitHub Actions and Airflow, automating model retraining and cutting manual release effort by 45% across 3 teams.
- Authored technical documentation and runbooks for model serving architectures, partnering with distributed product and engineering teams across US and Latin American time zones.

Software Engineer

Intellectsoft | New York, NY (Onsite) | Jun 2016 - Dec 2019

- Engineered Python and Java backend services and data pipelines for enterprise clients, processing 5M daily records and reducing API response time by **30%**.
- Implemented scikit-learn anomaly detection models for IoT and logistics platforms, lowering false alert rates by 25% and saving 15 support hours weekly.
- Collaborated with product managers and QA engineers on Git-based code reviews and CI workflows, improving release quality and documentation standards across client projects.

EDUCATION

Bachelor of Science in Computer Science

The University of Texas at Dallas | Richardson, TX | Aug 2015 - May 2019