**K.C. Tung, PhD**

Email: [kctung75034@gmail.com](mailto:kctung75034@gmail.com) Phone: (631)-943-4760 Citizenship: USA

Address: 5203 Lakehill Blvd, Frisco, TX 75034

Github: isaid22

***Proven GenAI Developer and thought leader with track records in delivering scalable solutions***

**HIGHLIGHTS**

* Proven leader in designing, deploying, and scaling AI/ML solutions across cloud platforms and distributed infrastructure in highly regulated enterprise environments.
* Deep expertise in LLMs, generative AI, and transformer models with hands-on optimization using XLA, TensorRT, and custom silicon accelerators.
* Strong track record integrating AI/ML into production systems using infrastructure-as-code, CI/CD, Kubernetes, and model monitoring frameworks.
* Trusted advisor to stakeholders across engineering, product, and compliance—bridging technical depth with clear communication and business alignment.
* Experienced in enforcing model governance, risk controls, and ethical AI principles for safe, auditable ML deployment at scale.
* Active thought leader: published author, speaker, and mentor in the AI community; contributor to TensorFlow ecosystem and enterprise GenAI enablement.

**CORE SKILLS**

* **Programming & ML**: Python, C++, Spark, TensorFlow, PyTorch, JAX, Ray, HuggingFace, Scikit-learn, CUDA, TensorRT, XLA.
* **AI/ML Ops**: FastAPI, Triton Inference Server, TensorFlow Serving, Docker, Kubernetes, CI/CD, MLflow.
* **Cloud Platforms**: Databriocks, AWS (SageMaker, Glue, DynamoDB, EventBridge, S3, Step Functions, Lambda, EKS, ECS), Azure (AzureML, AKS), GCP.
* **Data & DevOps**: Apache Kafka, Terraform, Airflow, SQL, NoSQL, Linux, GitHub, BitBucket, Jules, Spinnaker.
* **Specializations**: LLM, Generative AI, RAG, AIOps, Model Governance, Compliance, Network Automation.

**EXPERIENCE**

**JPMorgan Chase - CCB** Vice President, Applied AI/ML Lead 4/2024 – present

* Lead the design and deployment of generative AI applications in compliance-heavy production environments using LLMSuite API and infrastructure-as-code pipelines.
* Architect end-to-end ML solutions with progressive delivery strategies (A/B testing, canary deployments) for real-time campaign audience measurement, engagement optimization and lead optimization.
* Serve as an Application Owner (AO), define model governance, data usage controls, and guardrails to ensure safety, risk mitigation, compliance, and audit readiness in generative AI deployments.
* Collaborate cross-functionally with business and network engineering teams to automate decision-making workflows using ML.
* Mentor junior engineers, establish knowledge sharing, and drive experimentation culture across the firm’s applied AI teams.

**Amazon AWS** Senior Solutions Architect 7/2021 – 3/2024

* Designed and delivered AI/ML pipelines for enterprise clients across NLP, image generation, and time-series models using GPU/XLA, Neuron SDK, and AWS ML stack (SageMaker, ParallelCluster, Docker, EKS).
* Partnered with network and silicon engineering teams to benchmark model performance, optimize chip utilization, and provide insights from profiling tools like Nsight and Triton.
* Developed and deployed ML workflows using Kubernetes and CI/CD pipelines for scalable, resilient applications in production.
* Authored whitepapers, tutorials, and spoke at conferences to support adoption of AWS’s AI accelerators and distributed ML tooling.

**Microsoft Azure** Senior Cloud Solutions Architect 11/2018 – 07/2021

* Build and deploy deep learning ML models on customer data sets by utilizing distributed TensorFlow, Spark, AzureML services and Azure Kubernetes for scalable, cloud based serving.
* Collaborate with other WW AI Champs to draft and publish [***Cloud Adoption Framework – Machine Learning Security***](https://docs.microsoft.com/en-us/azure/cloud-adoption-framework/innovate/best-practices/ml-security) best practice for machine learning security in enterprise settings.
* Delivered scalable ML solutions for financial, manufacturing, and healthcare clients using AzureML, distributed TensorFlow, AKS, and Spark.
* Led secure ML architecture design aligned with enterprise data governance policies; co-authored Cloud Adoption Framework for ML security.
* Designed PoCs with real-time data ingestion, model inference pipelines, and hands-on customer enablement workshops.
* Advised data scientists and engineers on selecting scalable tools and best practices for cloud-based ML model lifecycle.

**AT&T** Principal Data Scientist 03/2014 – 10/2018

* Adopted deep learning (LSTM, CNN) techniques as an innovative solution to solve customer touchpoint prediction problem and enable data driven conversion funnel as a service for advertisers.
* Established best-practice strategy for data scientists, engineers and process owners to design scalable and production grade solutions in agile development environment.
* Translated user stories and business requirements to hypothesis for AI/ML models and designed end-to-end model solution architecture.
* Collaborated with industry thought leaders to evaluate data science platforms and vendors.
* Led, mentored, and developed strong data science and engineering talent in cross functional team environment.

**EDUCATION**

**Ph.D., Biophysics**, University of Texas Southwestern Medical Center, Dallas TX. *Dissertation: Human brain's neural network and synchronization activities (*[*https://pubmed.ncbi.nlm.nih.gov/23583747/*](https://pubmed.ncbi.nlm.nih.gov/23583747/)*)*

**B.S., Electrical Engineering**, University of Texas at Arlington, Arlington TX.

**CERTIFICATION**

* Microsoft Azure Certified AI Engineer, Data Engineer, Database Administrator, AI Ambassador.
* Microsoft Certified Enterprise Data Scientist.
* Microsoft OpenHack Coach: Knowledge Mining, Modern Data Warehousing, DevOps for Data Science.
* Microsoft GSMO Manufacturing EOU's AI Champ.
* [Microsoft Research AI School AI-611 Project MIKROS](http://aischool/Home/ai611.html) team member

**BOOKS and PUBLIC BLOGS**

* [***Learn TensorFlow Enterprise with Google Cloud AI Platform.***](https://www.amazon.com/Learn-TensorFlow-Enterprise-workloads-seamlessly/dp/1800209142/ref=sr_1_4?crid=30DVSBK184ISJ&dchild=1&keywords=learn+tensorflow+enterprise&qid=1614453728&sprefix=learn+tensorflow+%2Caps%2C174&sr=8-4) Packt Publishing, 2020.
* [***TensorFlow 2 Pocket Reference***](https://www.amazon.com/gp/product/1492089184/ref=dbs_a_def_rwt_bibl_vppi_i1)***.*** O’Reilly Publishing, 2021.
* [***Research and develop novel techniques to measure machine learning training data similarity.***](https://pureai.com/articles/2021/04/08/similarity-algorithm.aspx) Pure AI, 2021
* [***Scaling Large Language Model (LLM) training with Amazon EC2 Trn1 UltraCluster.***](https://aws.amazon.com/blogs/machine-learning/scaling-large-language-model-llm-training-with-amazon-ec2-trn1-ultraclusters/) 2023
* [**Maximize Stable Diffusion performance and lower inference costs with AWS Inferentia2.**](https://aws.amazon.com/blogs/machine-learning/maximize-stable-diffusion-performance-and-lower-inference-costs-with-aws-inferentia2/)2023
* [***Optimize AWS Inferentia utilization with FastAPI and PyTorch models on Amazon EC2 Inf1 & Inf2 instances.***](https://aws.amazon.com/blogs/machine-learning/optimize-aws-inferentia-utilization-with-fastapi-and-pytorch-models-on-amazon-ec2-inf1-inf2-instances/) 2023

**CONFERENCES AND WORKSHOPS PRESENTATIONS**

* **AWS Re:Invent 2022, 2023 workshop:** [**Stable Diffusion and Vision Transformers deployment on AWS**](https://catalog.us-east-1.prod.workshops.aws/workshops/c93ada51-f6db-4c36-a88c-47cc47e1cc60/en-US)
* [**A novel solution for data augmentation in NLP using TensorFlow**](https://www.oreilly.com/library/view/oreilly-tensorflow-world/0636920333111/video328030.html)**.** TensorFlow World. 2019. Santa Clara.
* [**A novel adoption of LSTM in customer touchpoint prediction problems**](https://conferences.oreilly.com/artificial-intelligence/ai-ca-2018/public/schedule/speaker/285298.html)**.** O’Reilly AI Conference. 2018. San Francisco.

**MY GITHUB CONTRIBUTION**

* Main page: https://github.com/isaid22
* Book: <https://github.com/PacktPublishing/Learn-TensorFlow-Enterprise>
* Container example: <https://github.com/isaid22/Tensorflow-Neuronx-Dockerfile>
* CUDA examples: [[link to CUDA examples]](https://github.com/isaid22/CUDA-examples)
* Nvidia Nsight: [[link to profiling a model in GPU]](https://github.com/isaid22/Profiling-TensorRT-Model-with-Nvidia-Nsight-Systems)
* Nvidia Nsight setup:[[link to Nsight setup]](https://github.com/isaid22/Nvidia-Nsight-Systems-Setup)
* Nvidia Triton Server: [[link to Triton Server setup]](https://github.com/isaid22/Triton-Server-on-Inferentia)
* Training LLM with Slurm Cluster using AWS ParallelCluster: [[link to setup]](https://github.com/aws-neuron/aws-neuron-parallelcluster-samples)

**PATENT**

* [**Velocity-weighted analysis of user equipment location data**](http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PTXT&s1=%22velocity+weighted%22&s2=10097960.PN.&OS=)- US10097960, Oct 9, 2018