Jules M.

Julesmhad@gmail.com | Website | (901) 648-2857 | LinkedIn | San Francisco CA

6+ yrs building high throughputs and low latency systems and models spanning **Ads Retrieval**, **NLP**, **GNNs**, Large Scale Recommenders.
Frontend:- Sveltekit, React, Typescript. **Backend:**- C++ Go GraphQL **Python,** Scala; **Spark**, **Kafka**, **Azure**, AWS and GCP cloud services

RELEVANT EXPERIENCE

Contra, **Software Engineer,** San Francisco CA — May 2023 – 2024

- Built large-scale traffic/observability stack: deployed an Envoy-based mesh with xDS control plane, canary + shadow traffic, and per-tenant rate limiting—handled 30M+ RPS at 99.99% availability while lowering p99 latency from 220ms → 95ms (−57%); implemented OpenTelemetry pipelines (logs/traces/metrics ~150TB/day) that cut MTTD by 40% and MTTR by 55%.
- Architected and deployed a distributed inference cluster (8 × A100 GPUs) on AWS EKS using Terraform and Helm, scaling to 2 k QPS at < 40 ms p99 latency, reducing per-request inference cost by 30% ($0.018→$0.012) and boosting system uptime to 99.95%.

Microsoft, Applied Scientist II, Bellevue WA — Aug 2021 – Mar 2023

Embedding based Ads retrieval at Audience Intelligence Platform Team

- Spearheaded the rollout of three GPT-driven ad query pipelines for 100 M+ Bing users—serving 2 k QPS with < 50 ms p99 latency—boosting ad match rate **by 25 %, CTR by 20** %, and generating an incremental $xM in annualized revenue, while streamlining MLOps to cut deployment cycles from 4 weeks to 3 days and rollback incidents by **60 %.**
- Built multi-task GNN embedding models serving **2B** global users, **1B Microsoft** users & **50B events** to personalize/ target ads for CVR-CTR prediction tasks with user behaviors, Ads features + Ads serving, User+Ads ranking service, DGL, DeepGNN, SQL, Java, **Spark, Kafka**

Income Targeting Product & Microsoft Shopping Team

- Spearheaded and launched Bing Ads' household-income targeting pilot—merging label-proportion learning, hybrid ZIP fallback, CLM bias correction, and mixed-effects intercepts—within a PyTorch mini-batch KL-divergence pipeline processing **50 M+** monthly impressions; delivered **85 % precision** in the Top 20 % income segment, **+22 % CTR** lift for finance campaigns, **sub-100 ms** p99 inference (< 70 % latency), and real-time bid adjustments that **boosted ROI by 18 %** on Top 10 % income audiences
- Deployed hybrid ZIP-level fallback to further extend coverage in low-traffic ZIPs by **40 %**, improving recall for income-segment inference by **25 %** versus pure ZIP-based targeting across all income buckets (Top 10 %, 11–20 %, … Lower 50 %)

Profile Prediction Model for Multilingual and Multi-country Markets

- Developed on, and enhanced Bing Ads' age-inference ensemble—merging an SVM "Age v1" classifier with a multi-task GNN "Age v4" embedding model (DGL/DeepGNN), ingesting 500 M+ daily behavior events and Microsoft Graph signals to infer missing user ages; integrated wide (numerical) and deep (text) features to boost age-prediction accuracy by **3 %**, cut mis-targeted ad impressions by **4 %**, and drive an **8 % CTR lift** for age-segmented campaigns, with daily profile scoring for **200 M+** users.

IBM, Software Developer, Poughkeepsie, NY — Oct 2020 – Sep 2021

Hyper Protect Data Controller at Z/OS Performance Team, IBM Z,

- Spearheaded CI/CD–driven **performance regression tests** for HPDC via Jenkins and Gatling, automatically validating end-to-end encryption and masking across **25 TB** of synthetic PII datasets nightly and alerting on any sub-5% throughput degradation.
- Led the v-team to develop elastic autoscaling strategies in Kubernetes for HPDC pods based on Prometheus metrics (CPU, memory, QPS), reducing cloud compute costs by **30%** and ensuring sustained **99.9% SLA** compliance under variable workloads.
- Built a real-time telemetry pipeline for HPDC with OpenTelemetry, Apache Kafka, and Grafana, enabling sub-5-minute detection of encryption errors and latency spikes and driving a **40% reduction** in incident MTTR.

BEDC Electric Plc, **Software Engineer,** Benin, NG — Nov 2016 – June 2020

Customer Data Infrastructure at Platform Team

- Spearheaded our engineering team to develop end-to-end near-real-time and batch ETL pipelines in Python and PySpark on Kubernetes—ingesting **100 K events/sec** for **5 million customers** across **27 districts** into AWS S3, orchestrated via Amazon Kinesis and SNS/SQS, with CloudWatch alerts and Kubernetes liveness probes ensuring **99.95 % uptime** for downstream energy-billing analytics
- Built and launched a Django REST API paired with a React/Chart.js dashboard on PostgreSQL—used by **200+ field engineers** to monitor daily usage deltas, resolve **1 500+ tickets/month**, and perform B2C triage, cutting average resolution time by **40 %**
- Spearheaded three cross-functional pods (10 engineers, 3 product owners) to pilot two personalized pricing offers for high-demand accounts—developing a Python/XGBoost ranking model that predicts prepaid quota breaches with **90 % precision**, driving a **20 % uplift** in top-up conversions
- Developed a day-ahead load-forecasting model by ingesting SCADA historian data via OSIsoft PI and implementing an LSTM network in PySpark—improving MAPE from 6 % to 2.5 %, saving **$1 M+** annually in reserve procurement costs
- Delivered an AWS Lex–powered chatbot integrated via Lambda on the customer portal—handling **300 K+** monthly interactions, deflecting **25 %** of support calls, and storing session data in DynamoDB before migrating to Couchbase for sub-50 ms read performance
- **Analyzed** multi-dimensional time-series data with PySpark and Pandas to engineer features for an LSTM-based anomaly detector, identifying **3 500+** high-reactance usage outliers per year and preventing **$500 K** in avoidable energy costs through automated alerts

EDUCATION & RESEARCH EXPERIENCE

Doctor of Philosophy, Computer Engineering, **University of Memphis** **inView**

Courses: Artificial Intelligence, Information Retrieval, Computer Vision, Image Processing, Data Mining, Deep Reinforcement Learning, NLP, NLU

Master of Science, Computer Engineering, **University of Memphis** **2020**

Courses: Advanced Algorithms, OOP, Web Mining & Search Engines, Machine Learning, DatabaseSystems, Adv. Statistics, Optimization

Proposed PySIM: a U-Net model for reconstructing 3D images from 2D layers captured from Structured Illuminated Microscopes

- Built TunableSIM GUI with C++ with Matlab's Engine API for **C/C++** and tested new features. Presented at OSI/COSI/SPIE Conference 2021
- 1st place in 2021 and 2020, at the University Research forum, two years in a row and regularly attended ML conferences.