# Chenru(Tyler) Lyu

25 Columbus Dr APT 5212, Jersey City, NJ 07302 cl3296@columbia.edu +1 929-602-3195

## OBJECTIVE

*To obtain a full-time senior level Software Development Engineer / Machine Learning Engineer / Data Engineer position*

## PROFESSIONAL EXPERIENCE

**Datadog,** *Senior Software Engineer in the Revenue Data Engineering Group, New York, NY,* *Apr 2023 – Present*

***Tech Lead*** *for multiple core data pipelines, both real-time and offline, and for machine learning model-based products*

- **Led Usage Forecasting and Anomaly Detection System** (collaborating with Machine Learning, Metering Platform, Revenue Query, and Frontend teams)
  *A new product for customer-facing usage forecasting and anomaly detection, as well as internal engineer-facing NRT alerts to promptly identify potential system issues*
  - **Lambda Architecture** (Metering Platform team)
    1. Fast Path: Developed NRT hourly usage forecasting and anomaly detection using an event-driven Luigi pipeline
    2. Slow Path: Developed daily usage forecasting and anomaly detection using a scheduled Luigi pipeline
  - **Data Quality Metrics Calculation and Kafka Event Driven** (Metering Platform team)
    1. Used Flink to consume real-time hourly and daily metering usage data points
    2. Conducted data quality analysis on each usage block dataset and published metrics to a Kafka topic
    3. Triggered fast path pipelines using a Kafka reader event loop to detect anomalies in usage data quality
  - **Model Training & A/B Testing** (Machine Learning team)
    1. Investigated various ML models, including Transformer and time series (SARIMA, Prophet)
    2. Launched model training and inference pipelines on a GPU Kubernetes cluster
    3. Utilized Metabase queries for backtesting and A/B testing to evaluate and iterate on models
    4. Supported a caching layer for model inference results per customer to enhance online prediction speed
  - **Model Serving** (Revenue Query team)
    1. Developed a backend model service for online usage prediction and forecasting
  - **Online Model Inference Platform** (Frontend team)
    1. Developed a frontend UI that integrates with the backend model service to perform inference and return results

**eBay,** *Technical Staff in the Recommender System Team, New York, NY,* *Dec 2019 – Apr 2023*

***Tech Lead*** *for machine learning-based item relevance rankers and ad revenue optimization in eBay's recommender system; led Agile development, new hire onboarding, and knowledge sharing*

- **Led Similar Item Deep Learning Ranker Model Project** (Team of 3 SDE, 1 Data Scientist)
  - Integrated BERT title embedding and ResNet image embedding into the new deep learning model
  - Developed a Spark ETL pipeline to extract daily incremental new item titles and load them into the HDFS directory
  - Created a Docker image for BERT model inference, scheduled as a daily job on the GPU cluster
  - Built a web application to fetch newly landed title embedding files from HDFS hourly and update Couchbase
- **Led Dynamic Ad Rate Weight Optimization (DARWO) Re-Ranker Project** (Team of 2 Data Scientist)
  - Compared XGBoost model-based re-ranker with a greedy algorithm re-ranker
- **Led Auction Promoted Listing (PLX) Ranker Project** (Team of 2 SDE)
  - Compared logistic regression model-based ranker with a heuristic scorer
- **Led Linear Programming Based Revenue Optimization Project** (1 Applied Researcher)
- **Presentation & Knowledge Sharing**
  - eBay Tech Blog: *An Introduction to Flink* Read here

**The New York Times,** *Senior Software Engineer in the Subscription Platform Team, New York, NY* *Jun 2019 – Dec 2019*

*Fully responsible for multiple core projects related to subscription backend services*

**Comcast/Freewheel,** *Software Engineer in the Linear Integration Team New York, NY* *Mar 2017 – Mar 2019*

*Fully responsible for the architecture design and final implementation of the creative service and transcoding service*

## SELECTED PROJECTS

**Augmented Luigi Orchestration (Python)** *Jun 2023 – Apr 2024 @* ***Datadog***

*The orchestration framework responsible for running all company-wide data pipelines using Luigi*

- Designed and implemented an enhanced Luigi framework for orchestrating data workflows, managing task dependencies, and ensuring efficient scheduling
- Implemented core Luigi components like scheduler and worker to manage parallel task execution and resource allocation
- Developed tasks with requires(), output(), run(), complete() methods to establish clear dependency chains, ensuring tasks were executed in the correct order
- Enabled on-demand task dependency generation in the run() function, allowing real-time updates to the DAG to reflect the evolving dependency chains
- Created a 'find best next task' algorithm in the scheduler to prioritize candidates through topological sorting based on DAG dependencies and submission order
- Implemented logging and monitoring in Luigi pipelines to track task status and optimize failure recovery and retry logic
- Designed a global lock mechanism for each scheduled Luigi task to ensure atomic task execution and avoid conflicts between pipelines running the same task

**Projected End-Of-Month Cost Forecasting & Cost Anomaly Detection Monitoring (Spark ETL/Scala/Python, data size 1 TB / day)** *Apr 2024 – Sep 2024 @* ***Datadog***

*A new product designed to help customers avoid unexpected high costs and identify specific products driving those costs*

- Collected the customer metering usage dataset, determined the projection window, and fed recent usage data into the extrapolated/ML model to produce the projected end-of-month (EOM) usage dataset
- Processed the projected EOM usage dataset through the monthly billing pipeline to calculate the projected EOM costs
- Stored the projected EOM cost dataset in S3 and made it accessible on Snowflake for backend services to ingest
- Built a DAG to orchestrate necessary tasks and utilized Terraform to schedule a daily Luigi pipeline on Kubernetes
- Collected customer behavior and time/product-dependent features, integrating them with the training dataset
- Scheduled a monthly model training pipeline for retraining XGBoost/transformer models on GPU to prevent degradation
- Developed EOM cost anomaly detection formula to calculate the normalized anomaly scores accross multiple granularities, such as billing dimension and datacenter, and established a reasonable threshold for detecting cost anomalies
- Built a Datadog monitoring system to auto-detect cost anomalies in real-time and send alerts to PagerDuty and slack

**Page-Optimization Transformer/RNN Model Pipeline** *Feb 2020 – Jun 2022* **@ *eBay***

*Achieved an 8% increase in monthly ad revenue compared to the previous month*

**Real-Time Feature Collection Data Pipeline – Poplup (Akka/Kafka/Scala, throughput 100k / s)**

*A real-time Akka stream pipeline to collect page-impression payloads and extract ML model features*

- Consumed Kafka page-impression topics, enriched, and combined features into Spark Rows
- Encoded enriched rows into local Parquet files, periodically syncing with HDFS

**Off-Line Feature Engineering Data Pipeline (Spark ETL/Scala, data size 1 TB / day)**

- Loaded page impression data in HDFS, attributing purchase/click labels to each placement row
- Mapped enriched rows into training feature values (feature mapping) and dumped them as NumPy files to HDFS

**Incremental Model Training Pipeline (Python/Docker/Jenkins)**

- Created a Docker image of the uWSGI web server for connecting to the Krylov GPU cluster to submit model training jobs
- Established a monthly Jenkins job to incorporate incremental training data and update the online model

**Event Tracer for Real-Time Feature Collection Pipeline (Akka/Scala/Html/JavaScript)** *Apr 2021 – Dec 2021* **@ *eBay***

*A full-stack project aimed at designing an event tracer for event visualization on the website*

**Back-End:**

- Developed a WebSocket event-driven connection handler to push event messages to the frontend in real time.
- Implemented fallback HTTP APIs for tracer session creation and event polling.
- Created an event publisher Akka actor to intercept useful intermediate events for debugging throughout the Akka graph.
- Designed a traffic sharding Akka actor to filter events exclusively for tracer-subscribed users.

**Front-End:**

- Established a JavaScript WebSocket connection with an onmessage function to collect event messages from the backend.
- Utilized Ajax as a fallback polling method to periodically call the backend event polling API for new event messages.
- Rendered an HTML page displaying new events received from the backend event polling API on the tracer UI page

**REST to gRPC Migration in Model Serving (Scala/Java/Springboot)** *Jun 2022 – Dec 2022* **@ *eBay***

*A Spring Boot-based gRPC application for online model serving*

- Defined gRPC services and implemented server-side methods, integrating gRPC client stubs
- Conducted load and performance testing at 20 TPS for both REST and gRPC endpoints
- Found gRPC to be 30% faster, with average response times of ~250ms for gRPC versus ~350ms for REST

**Akka Cluster Traffic Routing Architecture for Personalization Service (Akka/Scala/Springboot)**

*Jan 2022 – May 2022* **@ *eBay***

*A Spring Boot application for real-time personalization feature computation with traffic routing by UserID partitioning*

- Established an Akka cluster with a coordinator actor to manage individual worker lifecycles for computing user features, maintaining internal user feature state and recovering it from an external database on startup
- Utilized consistent hashing for coordinator registration, creating a group router for efficient message routing

**Flink Real-Time Streaming for Ultimately-Bought/Co-Sale Pipeline (Flink/Kafka/Scala, throughput 500k / s)**

*Jan 2020 – Dec 2020* **@ *eBay***

*A flink real-time pipeline to aggregate ultimately-bought and co-sale item pair counts within a recent 30-day time window*

- Built a Kafka source stage to consume View Item and Sale topics as a double-source stream
- Utilized a Keyed CoProcess operator to match ultimately bought pairs in a continuous session window with a 3-minute session gap, reflecting users' online shopping behavior
- Implemented an Async Sink stage to aggregate ultimately-bought/co-sale features and update Couchbase in near real-time

**Isotonic Regression Calibration for Similar Ranker (Python)** *Jan 2023 – Mar 2023* **@ *eBay***

- Created bins for the Similar Ranker model score range using quantiles and calculated the positive fraction in each bin
- Used sklearn package to fit model scores with an Isotonic Regression model and saved model parameters for online use
- Achieved improvement of PTR and NDCG through Isotonic Regression compared to Exponential and Sigmoid methods

**SKILLS**

---

**Programming Languages:** C/C++, Scala, Python, Java, Go, Ruby on Rails, Groovy, Unix Shell, SQL, Html/CSS/JavaScript, Swift, GLSL, R, MATLAB

**Cloud Infrastructure:** GCP (Google Cloud Storage, Pub/Sub), Azure (WASB, HDI), AWS (S3, EMR, SQS)

**Big Data Technologies:** Flink, Kafka, Spark, Hadoop, Hive, RabbitMQ, Flume, WANdisco Fusion

**Open Sources Technologies:** Springboot, Bootique, Akka, Mockito, Micrometer/Dropwizard/Prometheus, React/Angular/Vue, Cayenne, MyBatis, Ansible, Puppet, Oozie, Cucumber, OpenGL, OpenCV

**Software Tools:** Vim, Jenkins/Drone, Kubernetes, Docker, Harshicorp (Terraform/Vault), Jira/VSTS/TFS, GitLab/GitHub, Maven/Gradle/SBT, Bazel, Grafana, JMX, JMeter, VisualVM, Sumologic, Datadog, PagerDuty, Metabase, DbVisualizer, Lookers, Jupyter, Lenses, Presto, Kibana, Tmux

**Databases:** MySQL, Couchbase, Redis, Memcached, Elastic Search, Snowflake, MongoDB, Cassandra, HBase, DynamoDB, PostgreSQL, H2, Microsoft SQL, InfluxDB
**Function Skills:** OO Programming, Data Structure & Algorithm, Database Design, Web Analytics, Full Stack, CI/CD, Machine Learning, Neural Network, Statistical Analysis, Data Mining, Linear Models, Simulations

## EDUCATION

**Columbia University, New York, NY**
**M.S.**, Mechanics **GPA:** 3.6/4.0 *May 2016*
**Peking University, Beijing, China**
**B.S.**, Theoretical and Applied Mechanics **GPA:** 3.91/4.00 *Jul 2014*

## AWARDS AND ACTIVITIES

- First Prize, National Undergraduate Mathematical Modeling Contest, China *Jun 2013*
- First Prize, Undergraduate Physics Tournament, China *May 2012*
- Founder and Leader, Lead-guitarist, Columbia University Chinese Rock Band *Oct 2014 – Present*