

Dr. Zhao Tang

San Diego, CA | ztangintuit@gmail.com | linkedin.com/in/tzhao7-sde | github.com/zt5rice | zt5rice.github.io

SUMMARY

Software engineer focused on reliable data platforms, measurable AI products, and production systems. Two and a half years at Intuit owning real-time pipelines and customer-facing data products for TurboTax and QuickBooks (100M+ users), with experience across Spark, Databricks, ETL/ELT, data quality, on-call reliability, and AI-assisted data infrastructure. Recently built portfolio systems spanning streaming data, GraphRAG, and governed NL2SQL analytics. Strong advanced SQL and Python.

TECHNICAL SKILLS

SQL, ETL & Data Modeling: advanced SQL; ETL/ELT and data pipelines; schema design and query optimization; dimensional and normalized modeling; data warehousing; building foundational datasets from scratch; data quality metrics and validation

Data Platforms & Tools: Databricks, Delta Lake, lakehouse platforms, Spark, Airflow, dbt, Kafka, PostgreSQL, ClickHouse, Redis, AWS (EMR, S3, RDS, Lambda), GCP; Python (pandas), Docker, Kubernetes, Terraform, GitHub Actions CI/CD

Analytics & Measurement: experiment measurement and A/B analysis; metric and KPI definition; segmentation and customer journey analysis; dashboarding and self-service reporting

Languages: Python, SQL, Scala, TypeScript/JavaScript, Bash/Shell

Production & Collaboration: monitoring and alerting; on-call, production support, and root-cause analysis; stakeholder partnership; translating business questions into measurable outcomes

EXPERIENCE

Software Engineer — Customer Product & Experience Optimization

07/2025 – 07/2026

Intuit, Inc. | San Diego, CA — Customer Product Organization (CP) · TurboTax

- Defined data quality metrics for real-time pipelines and owned the fixes; built validation and monitoring that surfaced wrong data before it reached reporting.
- Upgraded batch pipelines to incremental processing, cutting peak processing time from 7 hours to 15 minutes and peak infrastructure cost by 95%.
- Led a cross-organization remediation of millions of missing records — framed the problem, drove agreement on one shared metric definition across three organizations, and set a single validation standard so every team could prove its part.
- Owned experiment measurement for TurboTax: raised device coverage in experiment data by 10+ percentage points (mid-60s → high-70s), making A/B readouts trustworthy enough to base product decisions on.
- Primary on-call engineer; used distributed tracing and root-cause analysis to cut MTTR by 40%.

Software Engineer — Data Platform & Analytics Infrastructure

03/2024 – 07/2025

Intuit, Inc. | San Diego, CA — A2D Analytics, AI & Data · TurboTax & QuickBooks

- Built real-time pipelines and foundational datasets for 100M+ TurboTax & QuickBooks users; used internal AI tooling to deliver 10+ pipelines with code review and end-to-end validation.
- Built a table-lineage capability mapping downstream dependencies across 100+ pipelines and ~1,000 tables — extracting read/write edges from each job's analyzed SQL in EMR Serverless logs plus CloudTrail events, then traversing the graph to produce an approval-ready view of affected tables, teams and managers. Cut manual cross-team approval coordination by ~50%.
- Fixed a multi-year stateful defect that reset visit counts annually — restored cumulative tracking since 2020 for 10M+ unique visitors and 1M+ daily sessions, stabilizing a metric behind multi-billion-dollar product decisions.

Research Specialist — Research Data Infrastructure (post-Ph.D.)

06/2023 – 03/2024

Rice University — Chemical & Biomolecular Engineering | Houston, TX

- Engineered reproducible multi-terabyte Python workflows with automated validation (>99% consistency); re-engineered the MATLAB tracking workflow into the open-source nanotrack pipeline supporting 6 peer-reviewed publications and additional posters/presentations.

Visiting Scholar — Computational Imaging & Machine Learning (post-Ph.D.)

05/2022 – 05/2023

Rice University — Chemical & Biomolecular Engineering | Houston, TX

- Built Python data-processing and machine-learning pipelines; parallelized processing across HPC clusters, cutting per-experiment runtime from 15 hours to 30 minutes.

DATA & AI SYSTEMS PROJECTS

clickstream — Real-Time Data Platform github.com/zt5rice/clickstream 2026

- Architected an end-to-end streaming data platform: Python producer (~100 events/s) → Kafka (KRaft) → Spark Structured Streaming (1-minute windows, watermarking, dead-letter topic) → PostgreSQL for curated aggregates + ClickHouse for analytics → read-only FastAPI.
- Built the identity, session and deduplication model downstream reporting depends on, plus automated data-quality checks, freshness SLIs, orchestration (Airflow/dbt), load testing (46 req/s, p95 181 ms, 0% errors) and 111 tests.

GraphAtlas — Multi-Engine GraphRAG Platform github.com/zt5rice/graphatlas 2026

- Built an end-to-end retrieval system: ingestion and entity/relation extraction, a PostgreSQL 17 + pgvector store with HNSW indexing, and retrieval fused from keyword + dense-vector + graph search with Reciprocal Rank Fusion, served through a TypeScript API.
- Engineered the evaluation harness behind it — a versioned 50-question golden set and a 4-mode ablation measuring hybrid Hit@5 of 0.840 vs 0.800 vector-only at p95 537 ms. Every metric traces to saved results, and chunk-level citations map each answer back to its source span.

NL2SQLAgent — Governed Natural-Language Analytics github.com/zt5rice/NL2SQLAgent 2026

- Built a self-service layer that lets non-technical users query a ~1.03M-row warehouse in natural language, with a defense-in-depth read-only SQL guard so no generated query can modify data.
- Tuned the query path against real data: GROUP BY/JOIN queries under 250 ms, end-to-end latency cut from ~25s to ~7s; 81 backend + 51 frontend tests.

nanotrack — Scientific Video-Tracking Infrastructure github.com/zt5rice/nanoparticle-video-pipeline 2026

- Re-engineered the MATLAB tracking methodology into an open-source Python package with NumPy, OpenCV, and scikit-image preprocessing, Dask parallelization, and checkpoint resume.
- Added FastAPI serving, an Airflow DAG, Docker Compose, Prometheus/Grafana observability, immutable run artifacts, tracking QC reports, and GitHub Actions CI.
- Parallelized the research workflow across HPC clusters, reducing per-experiment runtime from 15 hours to 30 minutes; the methods support 6 peer-reviewed publications and additional posters/presentations.

EDUCATION

Ph.D., Chemical & Biomolecular Engineering

Rice University, Houston, TX (May 2022) — computational imaging and machine learning; 6 peer-reviewed publications plus additional conference posters and presentations

D.B.A., Doctor of Business Administration — Business Analytics (in progress)

Westcliff University (May 2025 – present; expected Oct 2028) — doctoral coursework in research design & methodology: causal identification and threats to validity, quantitative research design and hypothesis testing, hierarchical regression and moderation analysis, reliability and validity testing (CFA, Cronbach's alpha, CR, AVE), SPSS

M.S., Computer Science (part-time, in progress)

Georgia Institute of Technology — expected 2027

B.S., Chemical Engineering (with honors)

Purdue University, West Lafayette, IN (2012)