Erica WU

Data Engineering | Software Engineer | Data Science

● Personal Website ● 📍 Los Gatos, CA ● 📧 xue.wu.erica@gmail.com ● 📱 +1-408-348-7931 ● U.S Citzen

SUMMARY

Data Engineer and Architect with 8+ years of experience building enterprise data models, large-scale batch and streaming platforms at Netflix, PlayStation, and Apple. Holds graduate degrees in Computer Science and Statistics, combining distributed systems expertise with statistical rigor. Translates complex marketing, financial, product, and commercial needs into end-to-end data solutions, from business requirements and data modeling to analytics, forecasting, and decision support. Designs governed data ontologies and semantic models that establish shared business meaning across systems, AI agents, models, and people. Leads AI-enabled engineering workflows with human-defined objectives and guardrails, auditable outputs, and clear human ownership of business decisions.

EXPERIENCE

Visionary AI — Data Architect & Software Engineer Data Platforms *Remote ● July 2024 ~*

- Designed and built a multi-agent analytics platform that evaluates content (scripts, video samples and IP) for IP procurement decisions, in which agents handle the engineering workload, I own the rules and the decisions.
- Architected an end-to-end evaluation pipeline (ingestion → compliance gate → script and media analysis → scoring → commercial forecasting → decision dashboard). It runs 6 critical, 5 supporting and 2 failure diagnose specialized AI agents, orchestrated with Dagster, on top of a bronze/silver/gold data model.
- Designed a 38-class 4-module (content, evaluation, commercial and governance) data ontology , with PROV-O lineage. It generates the agents' output schemas, the dbt tests and a shared glossary, so every project uses the same definitions. It contributes to our content library for IP and character reuse in future projects.
- Built deterministic scoring layers in dbt (6 dimensions, 24 sub-metrics, lint penalties, grade bands) driven by versioned rubric files to make scores reproducible and auditable, LLM agents extract evidence-backed features.
- Built a video analysis pipeline (shot segmentation, keyframes, speech-to-text, on-screen text, visual review). It produces a production-quality score kept separate from script quality.
- Developed revenue forecasting as P10/P50/P90 ranges, tagging every input with its source. A backtest loop against actual results to reduce forecast error.
- Established AI governance controls: least-privilege data access for agents, schema-validated agent outputs, golden set that catches regressions when prompts or models change and an independent agent that re-checks scores and reconciles numbers across reports human approval at 5 checkpoints.

Netflix Inc — Senior Data Engineer *Los Gatos & Los Angeles ● Apr 2022 - May 2024*

- Led architecture and ownership of Netflix's cross-domain financial analytics platform, unifying Content Finance, Talent Finance, and Revenue data models to support title economics, production and P&L analysis, revenue reporting, GTM strategy, financial planning, and executive reporting.
- Led data quality, observability and governance initiatives end to end; built governed semantic models with PII controls, ontology definitions, and certification workflows, and pipeline health scoring that improved data trust, detected pipeline failures earlier, and reduced recovery time for critical finance and content analytics pipelines.
- Crafted curated Spark batch pipelines processing 250M+ financial records and activity dataset; enabled standardized title-level economics, profitability analysis, finance close reporting, and content investment review.
- Designed Flink streaming pipelines for near real-time actual-spend tracking across 10M+ spend-related records, including PO invoices, payroll entries, journal entries, and GL postings; streamed curated actuals into Pinot and Druid to support cost-to-date reporting, actuals-vs-budget analysis, overrun-risk monitoring, accrual triggers, and re-forecasting workflows for production and content finance teams.

- Architected and optimized data storage layers across S3, datalake(Apache Iceberg), and Snowflake/BigQuery, designing data models, partitioning schemes, indexing strategies, and storage layouts to support product analysis, experiment and ML forecasting with cost-aware query performance.
- Built Vendor Contract Interactive legal Retrieval-Augmented Generation (RAG) systems combine LLM model with curated legal databases to let legal and operation teams query and explore contract information fast and safe.

Sony Interactive Entertainment (PlayStation) — Data Engineer *San Francisco ● Nov 2020 - Apr 2022*

- Built real-time Kafka and Flink streaming pipelines for PlayStation Store behavior and gameplay heartbeat telemetry, processing up to 5M events/sec (~7.5 GB/sec) and 3.23M events/sec (~4.0 GB/sec) respectively. Enriched events with user profile, ownership/entitlement, pricing, and promotion data to support next-item recommendation and homepage personalization models.
- Developed Spark batch pipelines for promotion schedules, campaign metadata, and campaign exposure datasets, processing 1–50 GB/month of promotion and campaign metadata and 2.5–150 TB/month of campaign exposure data. Modeled key campaign attributes including sale ID, discount, start/end time, campaign priority, creative ID, exposure channel, and exposure timestamp to generate personalized deals and store merchandising candidates.
- Designed scalable personalization data models and storage strategies, including partitioning, indexing, and optimized table layouts, improving query performance and reducing compute and storage costs across high-volume streaming and batch datasets.
- Created a data observability and quality platform for personalization pipelines, including dimension auditors, standardized error schemas, automated remediation workflows, Grafana dashboards, and backfill services to monitor data quality, recover failed jobs, and support historical reprocessing.

Apple Inc — Software Engineer, Core OS *Cupertino ● A*pr 2018 - Nov 2020

- Developed REST API backend for Xcode, serving 150K+ developers with performance and battery usage insights.
- Implemented data analytics pipelines integrating HDFS, Apple S3, and PostgreSQL, with analytical dashboards refreshing hourly to track OS functionality, battery drainage and device performance trends visualization and analysis for executive reviews.
- Processed telemetry from ~800M Apple devices worldwide into OLTP and OLAP systems (FoundationDB, SQLite) while ensuring compliance with regional privacy regulations.
- Designed data models for iOS/watchOS/macOS data, implemented OLAP cube, supporting 500+ reports and services used by 30+ teams.
- Built Tableau dashboards to surface OS system performance, power drainage trends and factors for VP-level executive review and diagnose OS performance and power issues.
- Built batch ETL pipelines for iOS 3rd-party app telemetry from 20M+ devices using Spark, Airflow, MongoDB, and PostgreSQL, ensuring high availability and scalability.

EDUCATION

University of California Davis

● M.S. Computer Science ● M.S. Statistics - Data Science

SKILLS

- **Programming:** Python, Scala, SQL, Java, Claude Code, AI Agents, dbt
- **Data Engineering:** Data modeling, Semantics, Ontology, ETL pipelines, feature engineering, vector embedding
- **Analytics & ML:** Regression analysis, clustering, multi-dimensional analytics, data & model benchmarking, RAG
- **Platforms & Storage:** Spark, Flink, Kafka, Airflow, Data Lake, Data Warehouse, Vector DB
- **Guardrails & Governance:** data quality, data governance, observability