Bo-Jian Ho

Data Engineer | AI Engineer
Sunnyvale, CA, USA, +1-626-662-5830, edwin.ho.bj@gmail.com

Professional summary

With 2 years of hands-on experience building scalable data platforms and production RAG systems. Skilled in designing lakehouse architectures, workflow orchestration, vector search and RAG pipelines. Focused on reliable data delivery and cost-efficient infrastructure, with a goal to advance data-driven products and enable intelligent, real-time analytics.

Employment history

Data Engineer, Aug 2025 - Present

Million, San Francisco, CA

- Architected scalable data lakehouse infrastructure using *Apache Iceberg* on *AWS S3*, *PlanetScale* (OLTP), and *ClickHouse* (OLAP), supporting **1TB+** daily data ingestion with real-time *Metabase* dashboards for operations teams.
- Engineered CDC pipelines using *Debezium* producer, *Kafka* as real-time streaming infra, and *ClickPipe* as consumer to migrate and sync **100TB+** historical data and **1TB+** daily data with zero downtime; reducing end-to-end data processing time by **70%**.
- Designed product usage data models that delivered critical insights across the full product lifecycle from development to deployment, with automated refresh processes via CDC ensuring real-time analytical data availability.
- Developed production-grade REST APIs handling **2B+** daily data point requests for operational analytics, optimizing query patterns, caching strategies to maintain sub-second response time; real-time cost dashboards resulting in **$2K** monthly savings.
- Built batched ETL workflows with *DBT* and *Dagster*, streamlining data flow via CDC and *Kafka* for faster business intelligence.

AI Engineer, Member of Technical Staff, Dec 2024 - Aug 2025

Chima, San Francisco, CA

- Developed end-to-end RAG system to enable semantic search for AI agents, architecting the full pipeline from query embedding generation using sentence transformers to vector similarity search with *Turbopuffer* and model-based reranking.
- Optimized with embedding caching, batch processing, hybrid retrieval combining semantic and keyword search using *LangChain* and *LangGraph* to achieve consistent sub-second response time while handling concurrent user queries at scale.
- Integrated semantic search capabilities with agent tools and context management, enabling agents to dynamically query knowledge bases, filter results by metadata, and chain multiple searches for complex information retrieval tasks.
- Implemented incremental updates that tracked file changes via *Git* and selectively re-embedded modified code sections, reducing full re-indexing time by **50%** while ensuring knowledge base always reflected the latest codebase state.
- Established evaluation using LLM-as-a-judge methodology to validate semantic search performance against traditional grep-based search, demonstrating a **70%** win rate in relevance and accuracy across diverse query patterns.

Data Engineer, May 2024 - Nov 2024

LESSO, Los Angeles, CA

- Developed forecasting models using *PyTorch*, *XGBoost*, and *LSTM* for demand prediction, combining time-series analysis with clustering techniques to identify seasonal patterns and improve inventory management decisions within **3 months**.
- Built automated ETL pipelines using *Python* and *SQL* to streamline data extraction on *AWS Glue*, reducing manual data preparation overhead and accelerating time-to-insight for business analysts.
- Created data model lineage visualizations enabling clear representation of complex data relationships and dependencies.
- Automated alerting system to detect inconsistencies in upstream tables and notify stakeholders of impacted data models and dashboards, enhancing data reliability and minimizing business disruption.
- Established data governance policies including role-based access controls and column-level masking in Snowflake to ensure compliance with data privacy requirements and secure handling of sensitive business information.

Student Engineer, Sep 2023 - Mar 2024

UC San Diego, San Diego, CA

- Optimized backend database queries through strategic indexing on frequently-accessed columns and query rewriting, achieving **~30%** improvement in average query execution time across production workloads.
- Implemented end-to-end ETL pipeline extracting data from AWS S3 using predefined schemas, applying data quality checks and business logic transformations including filtering and enrichment, then loading cleansed results into MySQL database.
- Developed shared Python library with reusable utilities and standardized patterns for common data operations, reducing code duplication and improving maintainability across the team.
- Established structured logging framework with contextual metadata to streamline debugging and operational monitoring.

Education

Bachelor of Science in Data Science, Sep 2021 - Mar 2024

University of California-San Diego, San Diego, CA

- GPA 3.8
- Minior in Cognitive Science
- Member of Data Science Student Society

Certifications

Certified Azure Data Engineer Associate

Microsoft Inc.

Advanced SQL for Data Scientists

DataCamp

Skills

SQL, Python, Typescript, Next.js, React, Distributed System, Machine Learning, PySpark, Apache Airflow, Apache Kafka, Apache Iceberg, DBT, PostgreSQL, MySQL, MongoDB, BigQuery, Data Modeling, Snowflake.

Links

LinkedIn: linkedin.com, GitHub: github.com, Personal Website: eddieho.xyz, Ami: ami.dev, Same: same.new, Langflow: langflow.org.