Yu-Chuan Hung

Data Engineer · Lakehouse, Batch & Streaming · Apache DataFusion Contributor

Email yuchuanhung.raymond@gmail.com | LinkedIn yuchuanhung | GitHub CuteChuanChuan | Blog raymondhung.dev

EXPERIENCE

Garmin
New Taipei City, Taiwan

Data Engineer I → II (promoted May 2026)
Jan. 2024 – Present

PLATFORM & DATA PRODUCTS

- *Lakehouse platform:* Designed and built a Kubernetes lakehouse on Spark and Iceberg, ingesting **900M+ records daily** into **500+ production tables** for application teams and production-line engineers; presented the architecture to US HQ
- *CDC streaming:* Built CDC ingestion with Kafka and Spark Structured Streaming into Iceberg for **70+ source tables**, delivering **sub-3-minute freshness**; idempotent `MERGE INTO` on primary keys keeps replays and retries safe
- *Test data:* Moved production test data to Iceberg, **cutting cross-version query time by about 90%**, enabling **sub-second product-version lookup across 700M+ rows** (previously infeasible) and faster data access for EE teams
- *Onboarding:* Designed YAML-driven pipeline generation, **cutting source onboarding from 6–8 h to 3 min**

RELIABILITY & OPERATIONS

- *Catalog migration:* Migrated the catalog layer from Hadoop Catalog to **Apache Polaris (Iceberg REST)** across **800+ tables**, eliminating commit failures from non-atomic S3 renames and enabling multi-engine access
- *CI/CD & monitoring:* Built CI/CD and orchestration on Argo Workflows/Events and GitHub Actions across **300+ pipelines** (**deployment time down 90%**), plus Grafana alerting on workflow failures and Kubernetes resource pressure

PERFORMANCE & COST

- *Spark shuffle:* Moved Spark-on-Kubernetes shuffle storage from NFS to local disk, **speeding up shuffle by 20–50%**
- *ETL rewrite:* Rewrote a business-critical legacy ETL with asyncio and Polars, **cutting runtime on 34M rows from 240 to 8 min** and reducing Oracle license cost

Academia Sinica & National Chengchi University
Taipei, Taiwan

Research Assistant
Aug. 2019 – Jul. 2023

- Built Python data-processing pipelines automating data cleaning and transformation, **cutting processing time by 60x**

PROJECTS

Apache DataFusion / arrow-rs | *Rust*
2025 – Present

- **10+ merged PRs** in DataFusion and arrow-rs across joins, kernels, Parquet, and Spark compatibility
- *Join execution:* fixed sort-merge join filter evaluation on NULL-padded rows (#26027), corrected `HashJoinExec` probe metrics (#25272), and refactored chunked probe-offset state into a typed enum (#26203)
- *Performance:* specialized `scatter` by type (#20498, **1.5–2.5x faster**), reused cached dictionary slots in aggregation (#25652)
- *Compatibility:* Added Spark-compatible `size` (#19592) and `json_tuple` (#20412); arrow-rs Bloom filter builder (#9877)

SKILLS

Languages	Python, SQL (Oracle, PostgreSQL, MySQL, MSSQL), Rust, Scala
Processing & Streaming	Spark (Batch & Structured Streaming), Kafka, Polars
Lakehouse & Query	Iceberg, Parquet, Polaris, Dremio
Infrastructure	Kubernetes, Docker, Argo Workflows/Events, Airflow, GitHub Actions, Grafana

EDUCATION

AppWorks School
Jul. 2023 – Nov. 2023

Data Engineering Program
Taipei, Taiwan

National Chiao Tung University (now NYCU)
Sep. 2011 – Jul. 2018

M.S. and B.S. in Industrial Engineering
Hsinchu, Taiwan