Chien-Hsiang Yeh

linkedin.com/in/chienhsiang-yeh | github.com/chypwc

Summary

Data Engineer with expertise in building scalable ETL pipelines, real-time inference systems, and AI-driven applications on AWS. Experienced in cloud-native tools (Glue, EMR, SageMaker, MWAA, Step Functions), data processing (PySpark, dbt, Airflow), and modern ML frameworks (LangChain, scikit-learn, RAG pipelines).

Skills

- **Languages:** Python, SQL, Bash, Yaml, STATA, MATLAB, JavaScript (Node.js), HTML/CSS
- **Databases:** PostgreSQL, MySQL, AWS RDS, DynamoDB, Qdrant, Pgvector, AWS S3 Vector Buckets
- **Data Warehouse:** Snowflake, AWS Redshift
- **AWS:** EC2, Glue, EMR, SageMaker, Lambda, Step Functions, Kinesis, MSK, MWAA, DMS, API Gateway
- **ETL Tools:** PySpark, dbt, Airflow, Kafka
- **API & Visualization:** FastAPI, Streamlit, Tableau
- **AI & ML:** Pandas, NumPy, scikit-learn, LangChain, Retrieval-Augmented Generation (RAG) pipelines,
- **DevOps & IaC:** Docker, Git, GitHub Actions, CloudFormation, Terraform

Experience

E-commerce ETL & RAG AI Agent github.com/InsightFlow8/insightflow

Description:

- Built an end-to-end ETL pipeline and AI-powered recommendation system for an e-commerce platform.
- The project automated data ingestion, transformation, and analytics to enable customer behaviour insights and personalised product recommendations.
- Agile practices followed with Jira and Confluence for sprint planning and documentation.

Key Responsibilities / Achievements:

- Designed and deployed ETL pipelines on AWS using Step Functions orchestrating Lambda, Glue Spark jobs, and Crawlers; implemented partitioned S3 Data Lake for scalable storage and Athena for analytics.
- Built a Streamlit dashboard (Docker + EC2) with FastAPI backend for real-time customer behaviour analysis, marketing insights, and AI-driven recommendations.
- Developed a RAG-based chatbot using LangChain and FastAPI; integrated ALS models and S3Vectors vector database for semantic product embeddings, similarity search, and personalised recommendations.
- Automated CI/CD pipelines with GitHub Actions and IaC with Terraform.

Real-Time Inference github.com/chypwc/kinesis-webui

- Engineered a serverless real-time recommendation system using Lambda, API Gateway, Kinesis, and DynamoDB for low-latency XGBoost predictions via RESTful APIs.
- Preprocessed data with Glue Spark and automated ML training and deployment on SageMaker, all orchestrated through Step Functions.
- Enabled global delivery with CloudFront; automated provisioning with Terraform and GitHub Actions.

DMS + EMR Data Pipeline github.com/chypwc/aws-dms-emr-terraform

- Automated ingestion and transformation from PostgreSQL into Apache Iceberg on S3 using AWS DMS, EMR Spark, and Glue Catalog.
- Orchestrated end-to-end processing with AWS MWAA (Airflow), Terraform-generated DAGs, and GitHub Actions CI/CD.

dbt–Glue ETL github.com/chypwc/aws-resources-exercises

- Implemented a scalable ETL pipeline by integrating Glue Jobs, Crawlers, and dbt-glue for SQL-based transforma-

tions; ingested Snowflake data into S3 (Parquet), cataloged with Glue, and transformed into features.

- Automated infrastructure and deployments using CloudFormation and GitHub Actions CI/CD, orchestrating ingestion, catalog updates, and dbt transformation runs end-to-end.

Research Assistant, ANU RSE — Canberra Oct 2024 – May 2025

- Structured raw datasets into panel format using STATA and Python.
- Built OCR pipelines to extract text from scanned documents.
- Conducted statistical analysis and literature reviews for labour economics projects.

Tutor, ANU RSE — Canberra Feb 2019 – Nov 2024

- Delivered tutorials and consultations across core economics subjects: mathematical methods, optimisation, macroeconomics, growth, and time-series forecasting (MATLAB).
- Taught econometrics with Python, focusing on statistical modelling and applied regression/causal inference.
- Led discussions on dynamic programming and general equilibrium; supported 500+ students across 20+ tutorials.

Education

Australian National University, Ph.D. in Economics July 2019 – July 2024

- Extended dynamic programming theory with state–action–dependent discounting.
- Proved convergence of Q-learning, SARSA, and Double Q-learning using real/functional analysis.
- Modeled equilibrium uniqueness in production and financial networks with Python simulations.

Australian National University, Master of Economics Feb 2017 – Dec 2018

National Tsing Hua University, B.S. in Physics; B.A. in Economics Sep 2008 – June 2012

Certificates & Compliance

- AWS Certified Data Engineer – Associate
- CPA Skills Assessment (Accountant – General)
- Working with Vulnerable People (WWVP)

Conferences

The Australasian Leadership Computing Symposium (ALCS) Canberra 2023

Presented: *Harold Zurcher as a Q-learner*

36th PhD Conference in Economics and Business Perth 2021

Presented: *Uniqueness of Equilibria in Interactive Network*

Society for the Advancement of Economic Theory (SAET) Canberra 2022

Presented: *Uniqueness of Equilibria in Interactive Network*

Referees

Available upon request.