Albert M. W. Yau

LinkedIn **GitHub** **Google Scholar**

Work Experience

R1 RCM | Senior Software Engineer (Data) 2023–Present

- Architected and implemented Spark Structured Streaming pipelines to replicate in-memory datastores to Databricks, processing 3M+ records/minute with < 60s p99 latency, reducing annual compute costs by $3.5M.
- Resolved concurrency bottlenecks of an AI charge categorization service by optimizing Spark DAGs to increase performance by 15x, and reconfiguring Spark clusters to reduce latency from 5 minutes to < 30 seconds.
- Designed and built a serverless data lake API, migrating web applications to query Databricks Delta Lake tables directly; eliminated a Redis caching layer based on load testing to simplify infrastructure.

dv01 (Acquired by Fitch Group) | Data Engineer (Infrastructure) 2018–2023

- Owned the end-to-end functional Scala/Spark data infrastructure for this Series A startup; architected reusable ingestion patterns to accelerate customer-facing integrations.
- Led a PB-scale storage migration from Azure to Google Cloud Platform; modernized legacy architecture to cloud-native services: SQL Server to BigQuery, Airflow to Cloud Composer, Spark to Dataproc, and VMs to Kubernetes.
- Executed a zero-downtime migration of a monolithic Scala codebase from Spark 2 to 3 ahead of Dataproc end-of-life, resolving cross-environment dependency conflicts across Python, Scala and R services.
- Prototyped from a Hack Day project, productionized a text parsing pipeline using custom Airflow operators and FastAPI microservices to extract unstructured PDF data; scaled capacity to 200+ securitizations.
- Designed and implemented the disaster recovery plan for financial data via multi-region replication and tiered backups to ensure high availability and data durability.

Stony Brook University | Graduate Student Researcher 2012–2018

- Provisioned and administered bare-metal Linux storage clusters (300+ TB capacity) using hardware RAIDs for high-throughput on-premises climate model analysis.
- Built data pipelines to aggregate 200 years of global climate model simulations across 30+ international institutions, standardizing datasets with different formats and spatial resolutions into a unified data store.
- Developed EOF-CCA statistical frameworks and storm track activity metrics to predict extreme weather patterns; publications were cited in the United Nations IPCC Sixth Assessment Report (AR6).

Technical Skills

Languages: Python, Scala, C#, C/C++, Fortran, MATLAB, Bash **Cloud:** Google Cloud Platform, Azure, Databricks
Data: SQL, Spark, Delta Lake, BigQuery **Frameworks:** NumPy, SciPy, Numba, Dask, Xarray, PyTorch, MPI, OpenMP
DevOps: Kubernetes, Docker, Terraform, Airflow **Systems:** Linux (Debian/Ubuntu/WSL), Proxmox, ZFS

Education

Stony Brook University MPhil, *Atmospheric Science* 2020
Scripps Institution of Oceanography, UC San Diego MS, *Oceanography* 2009
The Chinese University of Hong Kong BSc, *Theoretical Physics* 2008
Minors: *Computer Science*, *Mathematics*