ALBERT M.W. YAU

LinkedIn: albertmwyau
GitHub: mwyau

Data engineer experienced in managing end-to-end cloud data infrastructure, contributing to the successful exit of a fintech start-up. Previously a climate scientist with over 10 years of data analysis and high-performance computing (HPC) research experience, and an entrepreneur in the early 2000s.

SKILLS

Programming: Python, Scala, SQL, C#, RESTful APIs, Linux shell scripts.

Cloud Technologies: Google Cloud Platform, Microsoft Azure, Databricks, Cloudflare.

Data Engineering: ETL, Spark, Google BigQuery, Google Pub/Sub, Azure Service Bus, Airflow, MongoDB.

DevOps: Azure DevOps, GitHub Actions, Terraform, Kubernetes, Docker.

Certifications (Coursera Specializations): Data Engineering, Big Data, and Machine Learning on GCP; Machine Learning Engineering for Production (MLOps); Applied Data Science with Python; Deep Learning.

EXPERIENCE

Senior Data Engineer - CloudmedAI Platform, R1 RCM — Oct 2023 – present

Working within a team of full-stack software engineers to develop the next generation of AI applications, empowering healthcare revenue intelligence data:

- Took over an ongoing project to deliver within a short timeline:
  - Redefined the data lake schema, developed tooling for schema evolution, upgraded Databricks runtime and Spark, and optimized *Delta Lake* tables to unlock concurrent data processing to meet performance requirements.
  - Deep-dived into *Spark* DAG performance tuning, increasing data job performance by 15x.
- Proposed and executing a plan to migrate web applications from vendor databases to a serverless data lakehouse, aiming to reduce costs, increase reliability, and simplify the data pipeline.
- Collaborating with the QA engineer to create data integration tests to improve data quality.
- Collaborating with data scientists to establish the AI data ingestion pipeline.

Data Engineer - Infrastructure, dv01 / Fitch Solutions — Aug 2018 – Mar 2023

As one of the longest tenured data engineers, I was responsible for managing the end-to-end data infrastructure:

- Scaled the monthly mortgage-backed securities distribution day process through automation, ensuring hundreds of securitization reports and loan-level data to be processed within the same business day.
- Spearheaded the data pipeline Spark 3 migration as an advocate and the technical lead across the company, preventing *end-of-life* of old Google Cloud Dataproc images that would interrupt all production *ETL* jobs:
  - Upgraded the entire monolithic *Scala* codebase from 2.11 to 2.12.
  - Upgraded *Apache Spark* from 2.4 to 3.0 and upgraded all required dependencies, which is a fragile process.
  - Migrated data warehouse metadata from self-hosted *Apache Hive Metastore* to *Google Dataproc Metastore*.
  - Worked closely with data scientists to upgrade all production *R* Docker images to use the new data warehouse.
  - Managed Jira epics and tracked project progress; patched a conflict found on the deploy day to prevent a rollback.
- Acted as the technical lead of the data pipeline migration from *Microsoft Azure* to *Google Cloud* at the *petabyte* scale; executed a seamless transition with no data loss and zero downtime.
- Proposed the first continuous version update strategy of the data pipeline codebase; implemented automatic update alerts and scheduled security vulnerability checks through *GitHub Actions* and *Dependabot*.
- Initiated and implemented the *disaster recovery plan*; scheduled routine backups of the data warehouse to archive storage with limited access; validated the restoration process to protect business-critical data.
- Improved data warehouse load performance by 10x through migration from SQL Server to *Google BigQuery*.

- Deployed *Airflow* and maintained 1000s of DAGs; migrated self-hosted Airflow to *Google Cloud Composer*.
- Created, maintained, and migrated CI/CD pipelines in *Jenkins*, *Argo CI/CD*, *GitHub Actions* and *Terraform*.
- Created and maintained pipelines and backend services in *Scala* and *Python*, including automatic file scrapers and PDF parsers, *Airflow* plugins and operators, and API endpoints using *FastAPI/SQLAlchemy/PostgreSQL*.
- Worked in an agile environment, such as performing code reviews, sprint planning, and participating in hack days.
- dv01 was acquired by Fitch Group in September 2022.

Graduate Student Researcher, Stony Brook University Jun 2012 – Aug 2018

As an atmospheric science PhD candidate, I studied storm tracks using very large datasets from climate model simulations to diagnose long-term changes and impacts of extreme weather events:

- Administered Linux workstations and deployed new storage systems with over 300 Terabytes of local storage.
- Aggregated, processed, and managed a century's worth of climate data from over 30 global climate models in the CMIP5 archive, originating from 20 international institutions, on local storage systems. This streamlined access and analysis for researchers, contributing to multiple highly cited publications, with two cited in the *Sixth Assessment Report* (AR6) of the United Nations *Intergovernmental Panel on Climate Change* (IPCC).
- For my dissertation, I developed statistical frameworks based on principal component analysis (PCA) and canonical correlation analysis (CCA); the work was published in two parts in the *Journal of Climate* in 2020 and 2022.
- Investigated using *deep learning* to predict extreme weather events.
- Departed the position to join an early-stage Series A fintech start-up, *dv01*.

Data Science Intern, National Center for Atmospheric Research May 2015 – Jul 2015

I was accepted into the Summer Internships in Parallel Computational Science program at NCAR:

- Developed a Python-based cyclone tracker, *PyStormTracker*, over a ten-week period.
- Utilized *NumPy*, *SciPy* image processing library, and *MPI4Py* parallel programming library.
- Conducted benchmark tests on the Yellowstone supercomputer, with good scaling performance on 100s of CPUs.

HPC Support Assistant, The Chinese University of Hong Kong 2005 – 2008

I worked in the High-Performance Computing (HPC) Support Team under the Information Technology Services Centre to support the operations of on-campus supercomputing clusters:

- Assisted users including graduate students, postdoctoral researchers, and faculty members in using the HPC clusters:
  - Gathered requirements and installed dependent libraries.
  - Compiled, fine-tuned, and deployed scientific computing programs.
  - Debugged and troubleshooted programs on the HPC clusters, working closely with research users.
- Conducted *User Acceptance Testing* (UAT) for a newly constructed HPC cluster.

Founder, *XDDD*.org 2002 – 2005

I operated web hosting and webmail services in the early 2000s:

- Built a server which was hosted in the HKNet data center in Hong Kong.
- Implemented *LAMP* (Linux/Apache/MySQL/PHP) stack on RedHat Linux 9.
- Managed user dashboards, provided technical support, and responded to security incidents.
- Offered competitive pricing at HKD$100/100MB/year per user, with full Perl CGI and PHP support.

EDUCATION

MPhil Stony Brook University, New York *Marine and Atmospheric Science*

MS Scripps Institution of Oceanography, University of California, San Diego *Oceanography*

BSc The Chinese University of Hong Kong *Theoretical Physics*, minored in *Computer Science* and *Mathematics*