ALBERT M.W. YAU

GitHub: mwyau
LinkedIn: albertmwyau
Email: albert@mwyau.com
Google Scholar: V-2nZZ8AAAAJ

Data engineer with 5 years of experience in data and machine learning infrastructure, contributing to the growth of a successful fintech start-up. Previously a climate scientist with over 10 years of applied machine learning, Linux, and High-Performance Computing (HPC) research experience, and an entrepreneur in the early 2000s.

EXPERIENCE

Data Engineer, Infrastructure, dv01 / Fitch Solutions, New York, NY — Aug 2018 – Mar 2023

As one of the longest tenured engineers, I managed the end-to-end ETL data infrastructure and backend services:

- Spearheaded the data pipeline migration to Spark 3 as an advocate and the technical lead across all teams in 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 and reviewed PRs; patched an incompatibility found on the deploy day and prevented a rollback.
- Acted as the technical lead of the migration of the data warehouse 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 vulnerability checks through *GitHub Actions* and *Dependabot*.
- Put forward and executed the first *disaster recovery plan* in the company; scheduled routine backups of the *Apache Parquet* data warehouse and verified the restoration process to protect business-critical data.
- Improved data warehouse performance ten-fold through migration from Microsoft SQL Server to *Google BigQuery*.
- Maintained 1000s of active DAGs in *Airflow*; migrated Airflow from Compute Engine to *Google Cloud Composer*.
- Created, maintained, and migrated CI/CD pipelines in *Jenkins*, *Argo CI/CD*, *GitHub Actions* and *Terraform*.
- Created and maintained 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.
- Supported the machine learning pipeline in *Databricks* and *Airflow* utilizing the *H2O Sparkling Water* library.

Graduate Student Researcher, SoMAS / Stony Brook University, Stony Brook, NY — Jun 2012 – Aug 2018

As a PhD candidate, my expertise is in time series analysis; I studied storm tracks using very large climate model datasets:

- Administered workstations and deployed new storage systems with hundreds of terabytes of capacity.
- 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* (Convolutional Neural Network) 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, Boulder, CO — May 2015 – Jul 2015

I completed the Summer Internships in Parallel Computational Science program at NCAR:

- Over a 10-week period, developed a Python-based cyclone tracker called *PyStormTracker* from scratch.
- Utilized *NumPy*, *SciPy* image processing library and *MPI4Py* parallel programming library.
- Benchmarked on NCAR's Yellowstone supercomputer, with good scaling performance up to hundreds of CPUs.

HPC Support Assistant, The Chinese University of Hong Kong, 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) of 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 experience, provided technical support, and created dashboards; responded to security incidents.
- Offered competitive pricing at HKD$100/100MB/year per user, supporting up to 300 users with a 40GB hard drive.

Other Experience

Oceanographer at Coastal Environments (2010 – 2011); **Visiting Researcher** at Academia Sinica (2008) and Caltech (2007); **Summer Intern** at Hong Kong Observatory (2006); **Exchange Student** at University of Toronto (2006) and University of California, Berkeley (2005); **Network Technician** at Hong Kong Broadband Network (2002).

Skills

Programming: Python (NumPy/Pandas/PyTorch/FastAPI), Scala, Spark, SQL, RESTful API, Linux shell scripts.

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

Data Engineering/DevOps: ETL, Airflow, Hive, Google BigQuery/Cloud Run/Functions, GitHub Actions, Terraform.

Online Course Certifications: Deep Neural Networks with PyTorch, Natural Language Processing with Attention Models, Machine Learning Engineering for Production (MLOps) Specialization, Deep Learning Specialization, Applied Machine Learning in Python Specialization, Data Engineering on Google Cloud Platform Specialization.

Education

MPhil	School of Marine and Atmospheric Sciences, Stony Brook University, *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*

Other Interests and Hobbies

- Artificial intelligence: I used my NVIDIA GPU to train character-level RNN and nanoGPT with Chinese texts.
- Audio: I am fascinated by the physics of sound and enjoy reading objective reviews on Audio Science Review.
- Flightaware flight tracking: Raspberry Pi running PiAware with ADS-B antennas in the attic (feeder statistics).
- Home automation: whole home solar energy monitor using IoTaWatt / PVOutput, influxdb and Home Assistant; integrated Home Assistant dashboard showing personal weather station data, electrical loads, heating oil level, solar generation, water usage, and refrigerator/freezer temperature alarms.
- Home networking lab on a 10GbE Ubiquiti network stack:
  - NAS server with ZFS arrays, Raspberry Pi's, Jetson Nano, and 50+ IoT devices on an isolated VLAN.
  - Redundant Pi-hole DNS servers for family and ad blocking, using multiple DoH upstream for resiliency.
- Music: I enjoy a wide variety of genres, especially baroque music; I play the recorder (SSATB), flute, harmonica (beginner), highland bagpipe (practice chanter); I passed ABRSM Grade 5 in Descant Recorder and Music Theory.
- PC Building: I have been building my own computers and servers since I got first PC.
- Photography and astrophotography: I traveled to Oregon on Aug 21, 2017 to capture the total solar eclipse on video.
- Scouting: I achieved Challenger Award rank in Hong Kong; I enjoy hiking, camping, and traveling.