Albert M. W. Yau

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

Scientist-turned-engineer with 15+ years of combined experience bridging scientific research, high-performance computing, AI/ML, and petabyte-scale data infrastructure. Proven track record architecting streaming pipelines and data lakehouses, developing statistical frameworks for noisy real-world data, and scaling 0-to-1 production systems. Open-source contributor.

Experience

R1 RCM | Senior Data Engineer / Senior Software Engineer 2023–Present

- Architected and implemented custom CDC streaming workflows to replicate healthcare claims data to Azure Databricks, processing 3M+ records/minute with < 60s p99 latency and reducing cloud costs by $3.5M/year.
- Chose Spark Structured Streaming in Scala instead of Delta Live Tables, giving explicit control over scheduling, liquid clustering, checkpointing, and MERGE semantics that made the low-latency, low-cost design possible.
- Modeled medallion-style Delta Lake tables governed through Unity Catalog to support downstream analytics, AI workloads, and application APIs, with schema evolution, automatic backfills, and monitoring/alerting.
- Optimized Spark and MongoDB workflows supporting an AI categorization service by redesigning DAG execution and cluster configuration, improving throughput by 15x and reducing latency from 5 minutes to < 30 seconds.
- Designed a serverless data access layer for web applications to query Databricks Unity Catalog tables directly, simplifying architecture by eliminating Redis caching after validating performance through load testing.

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

- Owned the core Scala/Spark data platform for a Series A fintech startup, supporting petabyte-scale loan-level financial data; built reusable ingestion, normalization, and validation patterns for customer-facing integrations.
- Prototyped from a Hack Day project, then productionized a config-driven document processing pipeline using visual recognition, custom Airflow operators, and FastAPI microservices to extract unstructured PDF data; scaled capacity to 200+ securitizations.
- 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 large Scala codebase from Spark 2 to Spark 3 ahead of Dataproc end-of-life, resolving dependency conflicts across Python, Scala, R, and Databricks runtime environments.
- Designed and implemented a disaster recovery strategy for financial data using 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 with 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 extract signals from noisy climate data and predict extreme weather patterns; dissertation work published in Journal of Climate, with related publications cited in the United Nations IPCC Sixth Assessment Report.

National Center for Atmospheric Research | Data Science Intern 2015

- Built PyStormTracker, a high-performance cyclone tracker, within 10 weeks during NCAR Summer Internships in Parallel Computational Science (SIParCS).
- Accelerated cyclone detection using NumPy, SciPy, and scikit-image and scaled the tracking algorithm across 4,096 cores on the Yellowstone supercomputer using MPI4Py.

The Chinese University of Hong Kong | HPC Support Assistant / Student Researcher 2005–2008

- Optimized C/C++ and Fortran scientific workloads through compiler tuning and OpenMP/MPI parallelization.
- Deployed HPC clusters and built a campus-wide HDFS computational grid.
- Initialized and ran Weather Research and Forecasting model simulations on CUHK and Hong Kong Observatory clusters to model gravity waves over Hong Kong International Airport.

Projects

PyStormTracker | PyPI / conda-forge 2015–Present

- Originally developed for distributed HPC environments at NCAR using MPI4Py, with recent integrations of Dask multithreaded scheduling, additional tracking algorithms, and spectral/kinematics modules based on ducc0.
- Refactored core algorithms using vectorized NumPy arrays and Numba JIT-compiled kernels, with an 11x speed-up on serial workloads; results were compared with the legacy version to confirm a one-to-one match.
- Upgraded to Python 3.11+ with modern typing and linting, unit and integration test suite, and multi-arch/multi-platform Docker, PyPI and conda-forge publishing pipelines using GitHub Actions.

Homelab / Local AI Infrastructure 2019–Present

- Built a dual WAN 10GbE fiber network and Proxmox bare-metal virtualization environment with ZFS storage.
- Hosts containerized workloads and LLM experiments with local GPU inference.
- Manages IoT devices and contributed patches to open-source projects including IoTaWatt and graphs1090.

XDDD.org | Founder 2002–2005

- Built a web hosting and webmail service, with a physical server co-located in the HKNet data center.
- Created a secure multi-tenant Linux/Apache/MySQL/PHP (LAMP) stack on Red Hat Linux, with strict user isolation via chroot, suPHP, and disk quotas.

Skills

Languages: Python, Scala, SQL, C#, C/C++, Bash
Data Engineering: Spark, Spark Structured Streaming, Delta Lake, Unity Catalog, BigQuery, MongoDB, Airflow
Cloud / DevOps: Databricks, Azure, Google Cloud Platform, Kubernetes, Docker, Terraform, GitHub Actions
AI / ML / Scientific Computing: NumPy, SciPy, scikit-learn, PyTorch, JAX, Numba, Dask, Xarray, MPI, OpenMP
Certificates: Databricks Performance Optimization; Deep Learning / MLOps; Data Engineering on GCP

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
SURE Fellow: Caltech; Exchange: Toronto, UC Berkeley Minors: *Computer Science*, *Mathematics*