

# WILSON YIP

Data Engineer & Scientist bridging the gap between complex mathematical modeling and robust data infrastructure. Leveraged a background in Mathematics to optimize NLP algorithms and build high-frequency asynchronous scrapers. Expert in Python, dbt, and Airflow, with a focus on creating scalable, secure, and observable data environments that ensure data integrity.

## EXPERIENCE

### OneFineStay

Data Engineer

📍 London

📅 Oct, 2023 - Present

#### Data Engineering & Infrastructure

- Architected and maintained **ELT pipelines** using **Airflow**, **dbt**, **GCS**, and **BigQuery**.
- Developed **schema detection tools** to trigger full refreshes upon schema changes from upstream source.
- Optimized storage by implementing BigQuery-GCS External Tables, **eliminating data redundancy** and **enabling near real-time access**.
- Reduced query costs by implementing **Hive-partitioned directory structures** for external storage.
- Deployed CI/CD pipelines to automate testing on pull requests, **reducing production errors by 90%**.
- Engineered custom **dbt materializations** for BigQuery Functions to provide functionality ahead of native dbt-core support.

#### Data Observability & Cost Optimization

- Engineered a **cost-governance framework** by aggregating metadata from dbt manifest.json, BigQuery INFORMATION\_SCHEMA, GCP Audit Logs, and GCS Inventory Reports.
- Developed centralized observability tables to monitor tables, jobs, and GCS blobs, with automated reporting in Looker Studio.
- Reduced BigQuery expenditure by 80% through strategic partitioning, incremental modeling, query tuning, and storage billing optimization.

#### Cloud Infrastructure & Security

- Provisioned and managed GCP infrastructure using **Terraform** and **Docker**.
- Deployed Cloud Functions as webhooks for event-driven architecture.
- Implemented granular security protocols, including **column-level access control** and dataset-specific permissions.
- Containerized Airflow instances for scalable deployment to cloud services.

### Tailify Software

Junior Data Analyst (Machine Learning)

📍 London

📅 Jul, 2022 - Jul, 2023

- Engineered features and conducted EDA using **PySpark** and **ElasticSearch**, processing large-scale datasets to **improve model training quality**.
- Developed and deployed ML models to predict YouTube audience demographics, serving predictions via a high-performance **FastAPI backend**.
- Optimized NLP matching algorithms by introducing **soft-cosine similarity**, resulting in a **5–10% increase** in top-performer identification.
- Built **asynchronous URL scrapers** to resolve millions of shortened links, **reducing execution time by 90%** through concurrent processing.
- Architected and maintained PostgreSQL databases, collaborating with stakeholders to design schemas for **complex business requirements**.
- Orchestrated ETL pipelines using **Airflow** to ingest and transform agency performance and operational data.
- Implemented system observability by performing **log analysis** with **Grafana Loki** and building performance dashboards in **Grafana**.
- Accelerated internal workflows via rapid application development, automating document generation using **Google APIs**, **Slack API**, and **ElasticSearch**.

### Various Universities in Hong Kong

Research Assistant (Data Scientist)

📍 Hong Kong

📅 Sept, 2017 - Jan, 2022

- Perform statistical analysis and deploy machine learning models, including **AB-testing**, **PCA**, **Poisson regression**, **k-means**, **hierarchical clustering**, **LDA topic modelling**, etc. to perform analysis on different types of data. Developed and maintain **RShiny Dashboard** to visualise analysis results.

## EDUCATION

Society of Actuaries

Probability (P) Exam

University of Hong Kong

Bachelor of Science

Major: Mathematics/Physics

Minor: Computational and Financial Mathematics

## CONTACT INFO

in wilsonyip@elitemail.org

github https://github.com/wilsonkkyip

website https://wilsonkkyip.github.io

## SKILLS

### Highlights

Python	🐍	● ● ● ○
GCP	☁️	● ● ● ○
Docker	🐳	● ● ● ○
Terraform	terraform	● ● ● ○
GitHub Actions	github	● ● ● ○
AWS	aws	● ● ● ○
Rust	rust	● ● ● ○

### Data Processing

R	qr	● ● ● ○
Airflow	airflow	● ● ● ○
BigQuery	bigquery	● ● ● ○
dbt	dbt	● ● ● ○
Spark	spark	● ● ● ○
Looker Studio	looker	● ● ● ○
Grafana	grafana	● ● ● ○
PostgreSQL	postgresql	● ● ● ○
Elasticsearch	elasticsearch	● ● ● ○
Tensorflow	tensorflow	● ● ● ○

### Miscellaneous

Apps Script	script	● ● ● ○
Linux Bash	bash	● ● ● ○

### Administrative

Markdown	markdown	● ● ● ○
Latex	tex	● ● ● ○

## LANGUAGES

English: Fluent

Cantonese: Native

Mandarin: Fluent

## RESUME VERSIONS

Markdown

HTML

PDF