



## KỸ SƯ DỮ LIỆU

NGÔ TRÍ DƯƠNG

### MỤC TIÊU NGHỀ NGHIỆP

Tôi muốn chia sẻ kiến thức qua việc viết blog, tổ chức hội thảo kỹ thuật hoặc đóng góp tài liệu chính thức để thúc đẩy cộng đồng Data Engineering trong nước phát triển mạnh mẽ.

### THÔNG TIN CÁ NHÂN

11/05/1983

Hà Nội

thaolinh252512@gmail.com

0357223893

www.website.com

### HỌC VẤN

- Cơ sở dữ liệu và khai phá dữ liệu tại Đại học Khoa học Tự nhiên – ĐHQG Hà Nội - Kỹ thuật phần mềm tại Học viện Công nghệ Bưu chính Viễn thông

### KỸ NĂNG

- NoSQL (MongoDB, Cassandra)

- PostgreSQL

- Airflow

- dbt (Data Build Tool)

### KINH NGHIỆM LÀM VIỆC

- **CLOUD DATA ENGINEER** TẠI SMARTBANK (2022-2023)

+ THIẾT KẾ KIẾN TRÚC DỮ LIỆU TRÊN AWS VỚI S3, GLUE VÀ REDSHIFT

+ TỰ ĐỘNG HÓA VIỆC INGESTION LOG GIAO DỊCH NGÂN HÀNG TỬ HỆ THỐNG CORE

+ THAM GIA ĐÁNH GIÁ BẢO MẬT VÀ CHUẨN HÓA SCHEMA DỮ LIỆU TÀI CHÍNH

- **DATA ENGINEER** TẠI CÔNG TY DATAFLOW VIỆT NAM (2021-2023)

+ THIẾT KẾ VÀ TRIỂN KHAI PIPELINE ETL SỬ DỤNG  
APACHE AIRFLOW

## SỞ THÍCH

- Đọc sách

- Nấu ăn

+ TÍCH HỢP DỮ LIỆU TỪ NHIỀU NGUỒN NHƯ MYSQL,  
GOOGLE ANALYTICS VÀ API BÊN THỨ BA

+ TỐI ƯU HOÁ QUY TRÌNH XỬ LÝ BẰNG SPARK TRÊN  
GCP DATAPROC

## NGƯỜI GIỚI THIỆU

- Bà Phạm Thanh Hương (Data  
Operations Manager - Công ty  
ReportPro) - huong.pham@reportpro.vn  
- 0912111222

## DANH HIỆU VÀ GIẢI THƯỞNG

- **2023** - Bằng khen về tối ưu chi phí lưu trữ dữ liệu cloud –  
Công ty CloudData

- **2022** - Top 3 nhân viên cải tiến hệ thống ETL tại SmartRetail

- **2021** - Kỹ sư dữ liệu xuất sắc quý I tại Công ty DataBridge

- **2022** - Top nhân viên cải tiến năng suất truy vấn dữ liệu SQL  
tại ReportPro

- **2020** - Top 5 kỹ sư xử lý dữ liệu lớn trong cuộc thi nội bộ AI  
Challenge

## CHỨNG CHỈ

- **2020** - Google Cloud Professional Data Engineer Certificate

- **2022** - Big Data Engineering with Apache Spark – edX  
(BerkeleyX)

- **2022** - Data Engineering on Google Cloud Platform –  
Coursera

- **2023** - Designing Data-Intensive Applications – O'Reilly  
Certification Program

- **2022** - ETL and Data Pipelines with Shell, Airflow and Kafka – Coursera

## HOẠT ĐỘNG

### - Thành viên nhóm phát triển hệ thống phân tích log tại Công ty SecureLog (2021)

- + Xây dựng pipeline thu thập và phân tích log từ nhiều nguồn hệ thống.
- + Triển khai hệ thống cảnh báo bất thường dựa trên mô hình thống kê.
- + Sử dụng ELK stack để trực quan hóa dữ liệu log.

### - Tình nguyện viên hỗ trợ tổ chức Data Day tại Vietnam Data Community (2022)

- + Thiết lập demo pipeline dữ liệu thời gian thực bằng Kafka.
- + Hỗ trợ phần trình diễn luồng dữ liệu streaming giữa các node Spark.
- + Giới thiệu cấu trúc cơ bản của hệ thống Data Lake.

### - Thành viên nhóm tư vấn kiến trúc dữ liệu tại Công ty SmartData (2021 - 2022)

- + Đề xuất mô hình lưu trữ dữ liệu dạng columnar cho truy vấn phân tích nhanh.
- + Tham gia thiết kế cấu trúc bảng dữ liệu trong hệ thống warehouse.
- + Xây dựng best practice về đặt tên và phân vùng dữ liệu.

### - Cộng tác viên dự án dữ liệu mở tại OpenData Vietnam (2020)

- + Thu thập và chuẩn hóa dữ liệu dân số từ các tỉnh thành.

- + Tích hợp dữ liệu vào hệ thống cơ sở dữ liệu MongoDB phục vụ phân tích.

- + Tạo dashboard đơn giản để trực quan hóa dữ liệu bằng Plotly Dash.

#### **- Người viết blog kỹ thuật về hệ thống dữ liệu tại Blog cá nhân datapi.dev (2022 - nay)**

- + Chia sẻ kiến thức về thiết kế hệ thống ETL quy mô lớn.

- + So sánh các giải pháp lưu trữ dữ liệu: PostgreSQL, Redshift, BigQuery.

- + Viết hướng dẫn cấu hình Airflow và orchestration jobs theo lịch định kỳ.

## **DỰ ÁN**

#### **- Data pipeline xử lý log giao dịch hệ thống ngân hàng (Data Pipeline Engineer, BankDataTech) 2020**

Lưu trữ, phân tích và phát hiện bất thường từ log giao dịch ATM và mobile banking.

- + Sử dụng Fluentd thu thập log và gửi về Kafka

- + Phân tích real-time qua Spark Streaming

- + Lưu dữ liệu đã xử lý vào Elasticsearch và trực quan hoá bằng Kibana

#### **- ETL tự động hóa cho nền tảng tài chính (ETL Developer, FinTrack) 2021**

Tự động hoá quá trình tổng hợp báo cáo từ nhiều hệ thống tài chính nội bộ.

- + Viết ETL pipeline bằng Python và Pandas

- + Tích hợp với hệ thống email để gửi báo cáo tự động hàng tuần

- + Quản lý lịch ETL với crontab và giám sát bằng Prometheus

**- Realtime Analytics cho hệ thống bán lẻ toàn quốc (Big Data Engineer, Retail360) 2023**

Xây dựng hệ thống xử lý dữ liệu bán hàng theo thời gian thực để hỗ trợ ra quyết định tức thì.

- + Sử dụng Kafka để thu thập dữ liệu từ các chi nhánh toàn quốc
- + Xử lý dữ liệu streaming bằng Spark Structured Streaming
- + Gửi dữ liệu về Redshift và hiển thị trên dashboard BI

**- Data Lake cho hệ thống quản trị khách hàng (CRM) (Data Engineer, CRMPro) 2022**

Xây dựng nền tảng lưu trữ dữ liệu tập trung phục vụ phân tích hành vi khách hàng.

- + Tạo pipeline ingestion từ Salesforce, Google Ads và Facebook API
- + Lưu trữ dữ liệu theo mô hình phân vùng S3 Data Lake
- + Sử dụng Airflow để lập lịch và monitor luồng dữ liệu hàng ngày

**- Kiến trúc dữ liệu cho nền tảng giáo dục trực tuyến (Cloud Data Engineer, LearnHub) 2022**

Chuẩn hoá kiến trúc lưu trữ và xử lý dữ liệu học viên để phục vụ dashboard học tập.

- + Thiết kế hệ thống lưu trữ với BigQuery và Data Studio
- + Xây dựng Dataflow jobs để xử lý dữ liệu sự kiện học tập
- + Tạo luồng dữ liệu từ Firebase tới Google Cloud

