Vũ Hữu Sỹ - Big Data Engineer

Hà Nội
(+84) 865 294 163 | syvh.de@gmail.com

vu-huu-sy | vuhuusy

Học Vấn

Đại học Bách Khoa Hà Nội (HUST)
Chuyên ngành: Hệ thống thông tin quản lý

- Bằng: Giỏi (3.47/4.0)

Kinh Nghiệm

Big Data Engineer 07/2024 – 08/2025
Tập đoàn Công nghiệp - Viễn thông Quân đội (Viettel Group) *Cầu Giấy, Hà Nội*
Vai trò: Full Time

- Tham gia cắt chuyển, xây dựng, phát triển và tối ưu và các luồng nghiệp vụ từ Data Lake cũ lên nền tảng Lakehouse mới.
- Vận hành Data Lake, đảm bảo các luồng xử lý dữ liệu chính xác, đồng bộ, tổng hợp và chia sẻ dữ liệu từ/tới các hệ thống.
- Làm việc với các đội dự án để phát triển các luồng tổng hợp dữ liệu theo nghiệp vụ yêu cầu.
- Xây dựng luồng xử lý dữ liệu real-time sử dụng NiFi, Kafka, Spark ở thị trường Viettel Campuchia và Burundi.
- Thiết kế schema, partitioning bảng và áp dụng định dạng Parquet, ORC kết hợp với nén dữ liệu, giúp giảm 30% dung lượng lưu trữ và cải thiện hiệu suất truy vấn đáng kể.
- Hỗ trợ đào tạo thành viên mới, giới thiệu về tech stack, workflow và các tool nội bộ trong dự án, giúp rút ngắn 50% thời gian onboarding và tăng tốc độ hòa nhập vào công việc.

Thành tựu:

- Nhân viên xuất sắc tháng 06/2025
- Nhân viên xuất sắc tháng 03/2025

Data Engineer Intern – Viettel Digital Talent 04/2024 – 06/2024
Tập đoàn Công nghiệp - Viễn thông Quân đội (Viettel Group) *Cầu Giấy, Hà Nội*
Vai trò: Part Time

- Tham gia các buổi đào tạo sâu về chuyên môn Data Engineering, tìm hiểu các kiến trúc hệ thống dữ liệu lớn hiện đại.
- Đạt top 5 sinh viên xuất sắc nhất trong Giai Đoạn I của chương trình.

Data Engineer Intern 11/2023 – 02/2024
Công ty TNHH Giải pháp và Phân tích dữ liệu Insight Data *Nguyễn Tuân, Hà Nội*
Vai trò: Part Time

- Tham gia các buổi đào tạo về Data Warehouse doanh nghiệp.
- Xây dựng luồng xử lý dữ liệu với công cụ ODI và tạo báo cáo với Power BI.

Chứng Chỉ và Giải Thưởng

Databricks Certified Data Engineer Associate 07/2025 - 07/2027
- Link: https://shorturl.at/yiLoZ

Ambassador Giai Đoạn I của VDT 2024: Data Engineering 07/2024
- Link: https://shorturl.at/H63sc

Vô địch cuộc thi SQL CHAMPIONSHIP SEASON 2 - INDA 12/2023
- Link: https://shorturl.at/Ag8Ca

SQL (Advanced) Certificate - HackerRank 12/2023
- Link: https://shorturl.at/D19yC

Kỹ Năng

Ngôn ngữ lập trình: Python, SQL, Scala

Cơ sở dữ liệu: Oracle Database, MySQL

Containerization: Kubernetes, Docker

Big Data: Airflow, Hadoop, Spark, Kafka, Debezium, NiFi, Hive, Trino

ETL/ELT: ODI, dbt, Apache Hop, Pentaho

MLops: MLFlow, Feast

Khác: Linux, Git, Jenkins, Gitlab, Prometheus, Grafana, AWS (cơ bản)

Dự Án Cá Nhân

Scalable Real-time Fraud Detection Engine Built on Lakehouse Architecture
Tech stack: Kubernetes, Debezium, Flink, Spark, Kafka, MinIO, Hive, Airflow, Feast, MLflow, DataHub, Trino, Superset, Prometheus, Grafana

- Thiết kế và triển khai hệ thống phát hiện gian lận thời gian thực dựa trên kiến trúc Lakehouse sử dụng các công nghệ Big Data mã nguồn mở.
- Hỗ trợ xử lý batch và streaming tuân thủ ACID trên MinIO sử dụng Delta Lake.
- Xây dựng pipeline xử lý dữ liệu thời gian thực bằng Flink và Spark; tích hợp Feast làm Feature Store phục vụ làm giàu dữ liệu streaming.
- Triển khai mô hình Machine Learning bằng MLflow; gọi API predict trực tiếp dữ liệu streaming từ Flink.
- Trực quan hóa chỉ số gian lận và kết quả dự đoán bằng Superset kết hợp phân tích từ Trino.
- Tối ưu pipeline Flink đạt độ trễ 20–40ms mỗi message và recall mô hình >95% với XGBoost.
- GitHub: **github.com/vuhuusy/data-lakehouse-platform**

Data Lake Platform for Analytics
Tech stack: Hadoop, Hive, Trino, Kafka, Spark, NiFi, Superset, Docker

- Xây dựng hệ thống Data Lake hỗ trợ batch và streaming triển khai trên Docker.
- Tích hợp Hadoop + Hive cho lưu trữ, Trino cho truy vấn SQL, Kafka + NiFi để ingest dữ liệu.
- Tạo dashboard thời gian thực bằng Superset và Grafana để giám sát pipeline.
- GitHub: **github.com/vuhuusy/data-lake-platform**

eCommerce Data Warehouse System
Tech stack: ODI, Oracle Database, Power BI

- Phát triển Data Warehouse nhiều lớp với Staging, Star Schema và Data Mart cho từng phòng ban.
- Triển khai SCD Type 2 bằng IKM để theo dõi thay đổi lịch sử trong bảng Dim.
- Thiết kế load plan cho xử lý ETL song song để tối ưu hiệu năng.
- Kết nối các Data Mart với Power BI để tạo báo cáo và dashboard.
- GitHub: **github.com/vuhuusy/data-warehouse**

Khóa Học Liên Quan

IBM Data Engineering - Coursera: 9/2023 - 12/2023
NoSQL, Big Data, and Spark Foundations - Coursera : 11/2023 - 1/2024