Le Thanh Dat

lethanhdat.me@gmail.com | 0981729676 | github.com/ltdthanhdat

Education

Hanoi University of Science and Technology 9/2021 – 9/2025
Faculty of Mathematics and Informatics

Experience

Data Engineer, Amira Holdings JSC 5/2025 – Present

- Developed automated data crawling systems from multiple sources (LinkedIn, job sites, company websites) using reverse engineering, Scrapy, and Playwright.
- Processed unstructured data (IR reports, press releases, company organization charts) to extract structured information.
- Built sitemap classification algorithms to categorize company website pages (contact forms, homepage, service pages, etc.) for targeted data extraction.
- Built matching algorithms to find and connect company and user information from different data sources.
- Deployed large-scale data processing pipelines using Apache Spark with Arrow integration as compute engine, combined with Daft and Ray to reduce JVM overhead, and Delta Rust for ACID merges.
- Integrated Elasticsearch with Bulk API, custom analyzers, mappings, and real-time indexing from data pipelines for full-text search on company and article data.
- Built test data and validation frameworks to ensure data quality.
- Optimized Airflow scheduling through asset and dependency management to avoid resource waiting and balance runtime across multiple tasks.
- Built CI/CD pipelines for data engineering with code coverage and automated deployment.
- Deployed and managed GitOps workflows with ArgoCD to automatically deploy Kubernetes applications.
- Built BBQ Booking Agent App using LangGraph and React with multi-step booking UI (location/time/seat selection, guest info, confirmation), shadow DOM integration, calendar view, state management, and payment link integration.
- Deployed Knowledge Graph RAG system and SQL Agent to search and retrieve information from databases.

Data Engineer Intern, Educa Corp 1/2025 – 5/2025

- Deployed and configured a lab environment on a bare metal server, and managed services using k3s and Helm for system components.
- Set up monitoring, logging, and alerting for data infrastructure using Prometheus and Grafana, including creating and maintaining dashboards to track Oracle Tablespace memory usage.
- Implemented CI/CD pipelines using GitLab CI/CD for data workflows.
- Managed metadata using OpenMetadata to track data lineage and data sources.
- Built ELT pipelines using Talend and Apache Airflow to automate data processing workflows and synchronize CRM data (contacts, accounts, deals, tickets) from source systems to data warehouse, enabling sales and customer service reporting.
- Processed and integrated study logs from multiple sources (video sessions, exercises, mini tests) to analyze student learning behavior and performance metrics.
- Utilized Kafka and Debezium to enable real-time data synchronization between systems.
- Scheduled and orchestrated data pipelines with Airflow, ensuring stable data flow.
- Built semantic layer using Cube for unified data access and analytics across multiple data sources, and developed custom connection configurations.
- Configured Tailscale with ACLs to manage network access permissions and secure internal service connections.

Skills

Data Engineering Tools: Apache Spark, Daft, dbt, Talend, Kafka, Apache Airflow

Lakehouse: MinIO, Trino, Iceberg, Delta Table

Database: ClickHouse, Neo4j, PostgreSQL

AI: LangGraph, LightRAG, Knowledge Graph

DevOps & Monitoring: K8s, ArgoCD, Jenkins, Linux, Tailscale, Prometheus, Grafana

Cloud: Microsoft Fabric, AWS EC2, AWS Lambda, AWS S3

Web Development: Node.js, ReactJS, Tailwind CSS

Web Scraping: Playwright, Scrapy, Reverse Engineering, Browser Pool

Programming Languages: Python, JavaScript, TypeScript, Bash

Lê Thành Đạt

lethanhdat.me@gmail.com | 0981729676 | github.com/ltdthanhdat

Học vấn

Đại học Bách Khoa Hà Nội 9/2021 – 9/2025
Khoa Toán - Tin

Kinh nghiệm

Data Engineer, Amira Holdings JSC 5/2025 – Hiện tại

- Phát triển hệ thống crawl dữ liệu tự động từ nhiều nguồn khác nhau (LinkedIn, job sites, company websites) sử dụng reverse engineering, Scrapy, Playwright.
- Xử lý các dữ liệu không có cấu trúc (IR reports, press releases, company organization charts) để trích xuất thông tin có cấu trúc.
- Xây dựng thuật toán phân loại sitemap để phân loại các trang website của công ty (contact form, trang chủ, trang service, etc.) cho việc trích xuất dữ liệu có mục tiêu.
- Xây dựng thuật toán matching để tìm và kết nối thông tin công ty, người dùng từ các nguồn dữ liệu khác nhau.
- Triển khai pipeline xử lý dữ liệu quy mô lớn sử dụng Apache Spark với Arrow integration làm compute engine, kết hợp Daft và Ray để giảm overhead từ JVM, và Delta Rust cho ACID merges.
- Tích hợp Elasticsearch với Bulk API, custom analyzers, mappings, và real-time indexing từ data pipelines để tìm kiếm full-text trên dữ liệu companies, articles.
- Xây dựng framework test data và validation để đảm bảo chất lượng dữ liệu.
- Tối ưu lập lịch Airflow thông qua việc quản lý assets và dependencies để tránh chờ tài nguyên và cân bằng thời gian chạy giữa các task.
- Xây dựng các CI/CD pipeline cho data engineering với code coverage và tự động deployment.
- Triển khai và quản lý các GitOps workflow với ArgoCD để tự động deploy các ứng dụng Kubernetes.
- Xây dựng BBQ Booking Agent App sử dụng LangGraph và React với UI đặt bàn nhiều bước (chọn địa điểm/khung giờ/ghế, thông tin khách, xác nhận), tích hợp shadow DOM, calendar view, quản lý state, và tích hợp link thanh toán.
- Triển khai hệ thống Knowledge Graph RAG và SQL Agent để tìm kiếm và truy xuất thông tin từ database.

Thực tập sinh Data Engineer, Educa Corp 1/2025 – 5/2025

- Triển khai và cấu hình môi trường lab trên server bare metal, và quản lý dịch vụ bằng k3s và Helm cho các thành phần hệ thống.
- Thiết lập hệ thống giám sát, logging và alerting cho hạ tầng dữ liệu bằng Prometheus, Grafana, bao gồm việc tạo và duy trì dashboard để theo dõi bộ nhớ Oracle Tablespace.
- Triển khai CI/CD trên GitLab CI/CD cho các pipeline dữ liệu.
- Quản lý metadata dữ liệu với OpenMetadata để theo dõi data lineage và nguồn dữ liệu.
- Xây dựng pipeline ELT sử dụng Talend và Apache Airflow để tự động hóa quy trình xử lý dữ liệu và đồng bộ dữ liệu CRM (contacts, accounts, deals, tickets) từ hệ thống nguồn vào data warehouse, hỗ trợ báo cáo bán hàng và chăm sóc khách hàng.
- Xử lý và tích hợp dữ liệu study logs từ nhiều nguồn (video sessions, exercises, mini tests) để phân tích hành vi học tập và hiệu suất của học sinh.
- Sử dụng Kafka và Debezium để đồng bộ dữ liệu theo thời gian thực giữa các hệ thống.
- Lập lịch và điều phối pipeline dữ liệu với Airflow, đảm bảo luồng dữ liệu ổn định.
- Xây dựng semantic layer sử dụng Cube để truy cập và phân tích dữ liệu thống nhất từ nhiều nguồn dữ liệu khác nhau, và phát triển cấu hình kết nối tùy chỉnh.
- Cấu hình Tailscale với ACLs để quản lý quyền truy cập mạng và bảo mật kết nối các dịch vụ nội bộ.

Kỹ năng

Data Engineering Tools: Apache Spark, Daft, dbt, Talend, Kafka, Apache Airflow

Lakehouse: MinIO, Trino, Iceberg, Delta Table

Database: ClickHouse, Neo4j, PostgreSQL

AI: LangGraph, LightRAG, Knowledge Graph

DevOps & Monitoring: K8s, ArgoCD, Jenkins, Linux, Tailscale, Prometheus, Grafana

Cloud: Microsoft Fabric, AWS EC2, AWS Lambda, AWS S3

Web Development: Node.js, ReactJS, Tailwind CSS

Web Scraping: Playwright, Scrapy, Reverse Engineering, Browser Pool

Programming Languages: Python, JavaScript, TypeScript, Bash