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

- Crawl data automatically from multiple sources (LinkedIn, job sites, company websites) using reverse engineering, Scrapy, Playwright.
- Process unstructured data (IR reports, press releases, company organization charts) to extract information.
- Build matching algorithms to find and connect company and user information from different data sources.
- Deploy large-scale data processing pipelines using Apache Spark as compute engine, combined with Daft, Ray and Rust to improve processing performance and optimize RAM.
- Integrate Elasticsearch to index and search company and article data.
- Build test data and validation to ensure data quality.
- Optimize Airflow scheduling through asset and dependency management to avoid resource waiting and balance runtime when there are multiple tasks.
- Build CI/CD pipeline for data engineering with code coverage and automated deployment.
- Deploy and manage GitOps workflow with ArgoCD to automatically deploy Kubernetes applications.
- Build BBQ Booking Agent App using LangGraph and Agent Chat UI with React to interact with agents.
- Deploy 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.
- Deployed and managed services using Docker Swarm 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.
- Utilized Kafka and Debezium to enable real-time data synchronization between systems.
- Scheduled and orchestrated data pipelines with Airflow, ensuring stable data flow.
- Configured Cloudflare Zero Trust and Tailscale to secure network connections and manage internal access.

Skills

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

Lakehouse: MinIO, Trino, Iceberg, Delta

AI: LangGraph, LightRAG, Knowledge Graph

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

Cloud: AWS EC2, AWS Lambda, AWS S3

Web Development: Node.js, ReactJS, Tailwind CSS, WebAssembly

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

Programming Languages: Python, JavaScript, Bash, Rust, Scala

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

- Crawl dữ liệu tự động từ nhiều nguồn khác nhau (LinkedIn, job sites, company websites) sử dụng reverse engineering, Scrapy, Playwright.
- Xử lý các unstructured data như IR reports, press release, company organization chart để trích xuất thông tin.
- 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 làm compute engine, kết hợp với Daft, Ray và Rust để cải thiện hiệu suất xử lý và tối ưu RAM.
- Tích hợp Elasticsearch để index và tìm kiếm dữ liệu company, articles.
- Xây dựng 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 khi có nhiều task.
- Xây dựng CI/CD pipeline cho data engineering với code coverage và tự động deployment.
- Triển khai và quản lý GitOps workflow với ArgoCD để tự động deploy các ứng dụng Kubernetes.
- Xây dựng BBQ Booking Agent App bằng LangGraph và Agent Chat UI bằng React để tương tác với agent.
- Triển khai Knowledge Graph RAG system và SQL Agent để tìm kiếm và truy xuất thông tin từ cơ sở dữ liệu.

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

- Triển khai và cấu hình môi trường lab trên server bare metal.
- Triển khai và quản lý dịch vụ bằng Docker Swarm 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.
- 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.
- Cấu hình và triển khai Cloudflare Zero Trust và Tailscale để bảo mật kết nối mạng, quản lý quyền truy cập và kết nối an toàn cho các dịch vụ nội bộ.

Kỹ năng

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

Lakehouse: MinIO, Trino, Iceberg, Delta

AI: LangGraph, LightRAG, Knowledge Graph

DevOps & Monitoring: K8s, Docker Swarm, GitLab CI/CD, ArgoCD, Jenkins, Linux, OpenVPN, Tailscale, Cloudflare Zero Trust, Prometheus, Grafana

Web Development: Node.js, ReactJS, Tailwind CSS, WebAssembly

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

Programming Languages: Python, JavaScript, Bash, Rust, Scala