TRAN NHAT DUONG

DATA ENGINEER

+84 825 687 941
nhatduong01012005@gmail.com
Github
Linkedin

ABOUT ME

Data Engineering student proficient in Python, SQL, and Airflow, with complementary skills in Node.js Backend development. Understands Machine Learning concepts and data preparation needs. Seeking an internship opportunity to apply engineering principles to real-world data challenges.

EDUCATION

University of Science, VietNam National University HCM — *2023 - 2027 (expected)*
Bachelor of Computer Science
GPA: 3.74/4.0

TECHNOLOGY

Programing Language: Python, C++, SQL, JavaScript
ETL / Orchestration: Airflow, dbt
Cloud Platform: Microsoft Azure (ADLS Gen2, Databricks)
DevOps/Tools: Docker, Git, Linux.
Machine Learning: Scikit-learn

PROJECT

VietNamworks DE Pipeline — **Jan 2026 - Present**

Role: *Data Engineer*

Tech: *Python, dbt, Airflow, PostgresSQL (Neon), Docker, Azure*

Github: *VietNamworks_DE_Pipeline*

- Architected a Cloud-Native ELT pipeline on Azure Data Lake Gen2 (ADLS) and PostgreSQL, implementing Medallion Architecture (Raw-Silver-Gold) to ensure data quality and lineage.
- Engineered scalable data ingestion using Python (Pandas, adlfs) and Airflow, replacing legacy local storage with hierarchical cloud storage (HNS) for optimized big data processing.
- Orchestrated modular transformations with dbt Core, utilizing Jinja templating for reusable logic and automated testing (schema, referential integrity).
- Containerized the entire infrastructure (Airflow, Redis, Postgres) using Docker Compose, ensuring consistent environments from development to production.

Vietnamese-Chinese Corpus Pipeline — **Nov 2025 - Dev 2025**

Role: *Data Engineer, Core Developer*

Tech: *Python, PyTorch (LABSE), Vecallign*

Github: *Vie_Chn align pipeline*

- Engineered an automated ETL workflow to ingest, clean, and segment raw bilingual text (JSON) into high-quality parallel datasets (CSV) for machine translation training.
- Implemented LaBSE embeddings and the Vecalign algorithm to resolve complex sentence mismatches (1-N, N-1) based on vector cosine similarity.
- Integrated cross-platform hardware acceleration support (NVIDIA CUDA & Apple MPS), significantly reducing processing latency for large-scale text embedding.

CERTIFICATIONS IELTS: 6.5 (2022 - 2024)