TRAN NHAT DUONG

Data Engineer

+84 825 687 941 | nhatduong01012005@gmail.com | github.com/TrNhDuong_ | linkedin.com/in/trannhatduong

ABOUT ME

Third-year Computer Science student at VNUHCM – University of Science, seeking an **Intern / Fresher Data Engineer** role. Experienced in building end-to-end ELT pipelines on cloud infrastructure (Azure), orchestrating workflows with Airflow, and transforming data with dbt. Passionate about scalable data systems and continuously improving pipeline reliability.

EDUCATION

University of Science – VNUHCM, Bachelor of Computer Science *2023 – 2027 (expected)*

GPA: 3.74 / 4.0

TECHNICAL SKILLS

Languages: Python (proficient), SQL (advanced), C++, JavaScript

ETL / Pipeline: Apache Airflow, dbt Core, PySpark

Cloud: Microsoft Azure (ADLS Gen2, Databricks)

Databases: PostgreSQL, Azure Data Lake Gen2

DevOps / Tools: Docker, Git, Linux CLI, GitHub Actions

ML / Other: Scikit-learn, PyTorch (LaBSE embeddings)

PROJECTS

VietnamWorks Data Engineering Pipeline *Jan 2026 – Feb 2026*

Role: Data Engineer (Solo) **Tech:** Python, Apache Airflow, dbt Core, PostgreSQL (Neon), Docker, Azure ADLS Gen2

GitHub: github.com/TranNhatDuong/VietNamworks_DE_Pipeline

Designed and implemented a fully automated, scalable **End-to-End ELT pipeline** that crawls, ingests, and transforms job-market data from VietnamWorks into analysis-ready datasets.

- **Medallion Architecture:** Engineered a Raw → Silver → Gold layered pipeline using Azure Data Lake Storage Gen2 and PostgreSQL, ensuring clear data quality boundaries at each stage.
- **Scalable Cloud Ingestion:** Built a Python-based ingestion module (adlfs) integrated with Airflow DAGs; leveraged ADLS Hierarchical Namespace (HNS) to optimise big-data read/write performance.
- **Modular dbt Transformations:** Authored 15+ dbt models with incremental materialisation and data-quality tests (schema.yml), reducing transformation runtimes by ~30%.
- **Containerised Infrastructure:** Dockerised the entire stack (Airflow, Redis, PostgreSQL) with docker-compose, achieving zero-config Dev-to-Prod parity.
- **Outcome:** Pipeline processes ~5,000 job listings per run with end-to-end latency under 10 minutes.

Vietnamese–Chinese Parallel Corpus Pipeline *Nov 2025 – Dec 2025*

Role: Data Engineer & Core Developer **Tech:** Python, PyTorch, LaBSE (Google), Vecalign, CUDA / Apple MPS

GitHub: github.com/TranNhatDuong/Vie_Chn_align_pipeline

Built an **end-to-end NLP data pipeline** to clean, segment, and align raw Vietnamese–Chinese bilingual text (JSON) into high-quality parallel datasets (CSV) for downstream machine-translation training.

- **High-Accuracy Alignment:** Achieved >90% bilingual sentence-alignment accuracy by combining LaBSE multilingual embeddings with the Vecalign dynamic-programming algorithm, handling complex 1-N and N-1 mismatches via vector cosine similarity.
- **Cross-Platform GPU Acceleration:** Integrated automatic hardware detection (NVIDIA CUDA & Apple MPS) reducing large-batch embedding latency by ~55% compared to CPU baseline.

- **Robust Pre-processing:** Implemented configurable text-cleaning and sentence-segmentation modules that normalised noisy web-scraped data, improving downstream alignment precision by ~12%.
- **Outcome:** Delivered a reusable, config-driven pipeline capable of processing 100K+ sentence pairs per session.

CERTIFICATIONS

IELTS Academic 6.5 – British Council | Valid 2022 – 2024