

LU XUAN DUONG

DATA ENGINEER

☎ 0917 837 287 | ✉ xuanduong271104@gmail.com | 🔗 linkedin.com/in/xuanduongvn | 🌐 xuanduong2711 | 🌐 Portfolio

SUMMARY

Final-year Data Science student with hands-on experience building scalable ETL pipelines and processing large-scale data (1M+ records) using Apache Spark, Airflow, and Kafka. Proven ability to deliver production-grade data infrastructure in internship and personal project settings.

EDUCATION

VNUHCM – University of Science

Bachelor of Science in Data Science | GPA: 3.48/4.0

Ho Chi Minh City, Vietnam

2022 – 2026

TECHNICAL SKILLS

Core: Python, SQL, C++

Big Data: Spark, Airflow, Kafka, ODI ETL/ELT (LKM/IKM/CKM)

Databases: PostgreSQL, MySQL, Oracle 19c, MongoDB, SQL Server

DevOps: Docker/Compose, Git/GitHub, Linux (Ubuntu, Oraclelinux)

Analysis: Pandas, NumPy, Parquet, Streamlit, Selenium

WORK EXPERIENCE

Saigon Thuong Tin Commercial Joint Stock Bank (Sacombank)

Database Administrator (DBA) Intern

Ho Chi Minh City, Vietnam

Jan 2026 – Mar 2026

- Engineered and optimized high-performance ELT pipelines to migrate **1,000,000+ records** daily across Oracle-to-Oracle environments using **Oracle Data Integrator (ODI)**.
- Leveraged **Check Knowledge Modules (CKM)** to implement data quality firewalls, automatically detecting and filtering **2% anomalous data** to ensure warehouse integrity.

PROJECTS

Cinema Data Platform

Oct 2025

- Built an ETL pipeline processing **15,000+ daily records** from 80+ cinemas into an **MSSQL** data warehouse orchestrated by **Airflow + Spark**.
- Developed a **Streamlit** dashboard for ticket price analysis, surfacing insights that enabled up to **20% cost savings**.
- Tech:** Python, Spark, Airflow, MSSQL, Docker
- GitHub:** Cinema Data Platform

Real-Time CDC Data Pipeline

Feb 2026 – Mar 2026

- Built a real-time Change Data Capture pipeline syncing MySQL to MongoDB using **Debezium + Kafka + Spark**.
- Automated multi-table setup by dynamically fetching schemas and provisioning Kafka topics directly from MySQL metadata.
- Reduced system overhead via parallel processing and MongoDB bulk writes.
- Tech:** Python, Spark, Kafka, Debezium, MySQL, MongoDB, Docker
- GitHub:** Real-Time CDC Pipeline

LANGUAGES

Vietnamese: Native

English: B2