



# Vo Duc Tuan

## Data Engineer Intern

Graduated in Software Engineering from the Vietnam-Korea University of Information and Communication Technology and completed the Data Engineer program at FUNiX. I am seeking a Data Engineer Internship to apply my knowledge of SQL, Big Data, Airflow, AWS, and Data Warehousing in building and managing data pipelines, while gaining practical experience, strengthening ETL skills, and deploying data systems on cloud environments.

## Personal Information

0818741182

13/05/2003

voductuan1305@gmail.com

<https://github.com/VDTune>

Ngu Hanh Son District, Da Nang City, Vietnam

## Projects

### Traffic Safety Data Engineering & Analytics System | 08/2025 - 09/2025

<https://drive.google.com/drive/folders/1FVH9L0ivOz14MdzrOPlcz2Zpsv4GZhO5?usp=sharing>

**Objective:** Develop an end-to-end data analytics platform to process, store, and analyze large-scale US traffic accident data (2016–2023). The goal is to support data-driven decision-making and improve traffic safety insights.

**Technologies:** Airflow, Apache Spark, SQL Server, SSIS, SSAS, Power BI, Visual Studio.

#### Description:

- Designed and implemented a **Snowflake Schema** to model relationships between accidents, vehicles, and drivers.
- Built automated **ETL pipelines with Airflow, Spark** for data collecting and cleaning.
- Developed **OLAP Data Cubes using SSAS** to enable multidimensional analysis.
- Applied **SSIS** for data staging, and warehouse loading.
- Created **interactive Power BI dashboards** to visualize accident trends by time, location, severity, and contributing factors.
- Ensured data accuracy, integrity, and scalability across the entire pipeline.

### Integrating the Movie Data Warehouse into AWS Redshift | 08/2025 - 09/2025

**Objective:** Build and deploy a data warehouse for Netflix movie data, migrating from an on-premises system to AWS Redshift to leverage cloud-based big data processing capabilities.

**Technologies:** SQL Server, SSIS, T-SQL, AWS Redshift, Amazon SCT, Amazon S3

## Education

### Vietnam-Korea University of Information and Communication Technology

| 2021 - 2026

Software Engineer

GPA: 3.0/4.0

### FUNiX Online University |

11/2024 - 9/2025

#### Data Engineer Program

Database Systems

Data Engineering

Big Data with Spark

Data Engineering on AWS

# Skills

---

## Data Programming Languages

Python (pandas, PySpark), SQL, MongoDB Query Language (MQL) , Bash

## Distributed Data Processing

Apache Spark, Hadoop  
MapReduce, Apache Kafka  
Streams, AWS EMR (Elastic MapReduce)

## ETL / Orchestration

ETL with SSMS, Visual Studio (SSIS); Workflow orchestration with Apache Airflow (Astro, CLI, Docker Desktop)

## Data Warehousing Redshift

BigQuery, Snowflake; Schema design (star/snowflake, partitioning)

## Cloud & DevOps

AWS (S3, Glue, EMR, Lambda), GCP (BigQuery, Dataflow), Terraform/IaC, Docker & Kubernetes, CI/CD (GitHub Actions)

## Soft Skills

Problem-solving  
Effective Communication  
Collaboration & Teamwork  
Adaptability

# Certifications

---

## 01/2025

TOIEC - 755

## 09/2025

Data Engineer - FUNiX

## Description:

- Imported Netflix dataset from CSV files into **SQL Server Source**.
- Built an **on-premises Data Warehouse** with ETL processes using **SSIS**.
- Converted ETL workflows into **T-SQL stored procedures** for optimization.
- Created an **AWS Redshift cluster** and migrated schema/procedures using **Amazon SCT**.
- Transferred data from **SQL Server → S3 → Redshift** using the **COPY command**.

## Big Data Pipeline Deployment on Cloud | 07/2025 - 08/2025

<https://github.com/VDTune/cloud-data-pipeline.git>

**Objective:** Design and implement an automated data pipeline to collect, process, store, and analyze big data from cloud sources.

**Technologies:** Apache Airflow, Python, Spark, MongoDB, Google Drive API, Bash

## Description:

- Designed an **Airflow DAG** with tasks for start, end, branching, clearing, downloading, importing, Spark processing, and saving results.
- Integrated **Google Drive API** to fetch raw data into the system.
- Imported raw data into **MongoDB** using BashOperator with mongoimport.
- Built a **Spark job** to compute the number of answers per question and exported results to CSV.
- Re-imported Spark output into **MongoDB** for analytical purposes.

## Building a COVID-19 Data Collection System | 06/2025 - 07/2025

<https://github.com/VDTune/covid19-data-crawler.git>

**Objective:** Collect and process COVID-19 case data from online sources to support analysis and outbreak monitoring.

**Technologies:** Python, Scrapy, Regex, JSON

## Description:

- Developed a **Scrapy spider** to automatically extract new case counts and timestamps from websites.
- Processed data using **Regex** and text preprocessing techniques to standardize information.
- Implemented **pagination handling** to retrieve historical records.
- Stored data in **JSON format** for further analysis.
- Extracted and normalized case data by province/city for detailed insights.