Vu Thien Son | Data Engineer

Phone: (+84) 825-143-790 **Date of Birth:** 28-1-2004
Address: Cau Ong Lanh ward, Ho Chi Minh City
Email: vuthienson280104@gmail.com **Linkedin:** linkedin.com/in/sonvuuu28

Object

I am a fourth-year Information Technology student at Sai Gon University with a strong interest in Data Engineering, currently seeking an internship to work with Big Data and data pipelines.

Education

Sai Gon University ***2022 - Current***
Bachelor of Engineering in Information Technology

Language

Toeic Listening and Reading English Certificate: 900

Skill

Programming: Python, SQL, Java, Scala
Databases: MySQL, Cassandra
Big Data: Spark, Kafka, Flink
Data Engineering: ETL/ELT, Data Pipeline, Data Modeling (Star/Snowflake)
DevOps & Tools: Docker, Airflow, Git/GitHub, Grafana, Power BI

Personal Project

Real-time Recruitment Data Pipeline ***Nov 2025 - Dec 2025***
Kafka, Flink, Mysql, Grafana, Cassandra, Pyspark, Airflow **github.com**

- Built a data pipeline system to ingest data from a recruitment website, using Kafka as a central data hub with two processing flows.
- In the batch flow, stored raw data from Kafka topics in Cassandra (Data Lake) and performed ETL with PySpark, loading curated data into MySQL Data Warehouse, orchestrated by Airflow.
- In the real-time flow, processed Kafka streams using Apache Flink to transform data in real time and load results into MySQL, visualized results through Grafana dashboards.
- Dockerized the entire system and deployed it on a single virtual machine.

Big Data ETL Pipeline for Customer 360 Analytics ***Sep 2025 - Nov 2025***
Pyspark, MySQL, PowerBI, LM studio **github.com**

- Built data pipelines to transform OLTP data into OLAP output, using a Customer 360 framework to create a unified customer view across touchpoints.
- Modeled fact tables for contract activities and customer interactions, with dimension tables for channels, applications, time, and customer attributes.
- Developed batch and near real-time (mini-batch) pipelines to support BI reporting and ad-hoc SQL queries.
- Integrated a classification model to improve the accuracy of customer search input categorization, supporting better customer analysis and personalization.

Big Data ETL Pipeline for Customer 360 Analytics

Sep 2025 - Nov 2025

Pyspark, MySQL, PowerBI, LM studio

github.com

- Built data pipelines to transform OLTP data into OLAP output, using a Customer 360 framework to create a unified customer view across touchpoints.
- Modeled fact tables for contract activities and customer interactions, with dimension tables for channels, applications, time, and customer attributes.
- Developed batch and near real-time (mini-batch) pipelines to support BI reporting and ad-hoc SQL queries.
- Integrated a classification model to improve the accuracy of customer search input categorization, supporting better customer analysis and personalization.

Big Data ETL Pipeline for Customer 360 Analytics

Sep 2025 - Oct 2025

Python, PowerBI, n8n

github.com

- Built data pipelines to transform OLTP data into OLAP output, using a Customer 360 framework to create a unified customer view across touchpoints.
- Modeled fact tables for contract activities and customer interactions, with dimension tables for channels, applications, time, and customer attributes.
- Developed batch and near real-time (mini-batch) pipelines to support BI reporting and ad-hoc SQL queries.
- Integrated a classification model to improve the accuracy of customer search input categorization, supporting better customer analysis and personalization.