PHAM TAN ANH VU

(+84) 934 627 450
Ho Chi Minh, Viet Nam
anhvuphamtan@gmail.com
Github : anhvuphamtan

Highly enthusiastic in Data Science fields with a strong drive to acquire, transform and analyze data to generate impactful insights. I am actively pursuing an internship or entry-level position as a Data Engineer, wishing to grow more in the field. I am aspiring to learn new technologies, and eager to tackle real-world challenges.

EDUCATION

Ho Chi Minh University Of Science

***Major**: Data Science*

***GPA**: 3.24/4.0*

SKILLS

Programming languages: Python, Java, SQL
Data processing frameworks: Apache Spark, Hadoop, Kafka, Airflow
Cloud: AWS (S3, Redshift)
Visualization: Grafana
Technical environments: Docker, Terraform

CERTIFICATES

English: IELTS 7.5

PROJECTS

Stream Processing: Real-time Click Attribution and Dynamic E-commerce Insights

https://github.com/anhvuphamtan/Stream_processing

Objective :

- Design E-commerce real-time tracking solution using **First Click Attribution** to identify **checkout-driven clicks** with large request volumes.
- Provide actionable, near real-time insights into marketing impact by analyzing **checkout-driven clicks** from various sources (e.g., FB ads, TikTok ads). Empower stakeholders to optimize marketing and business strategies based on these insights.

Key features :

- Handle streaming data sources with Kafka and Spark Structured Streaming.
- Data storage with PostgreSQL.
- Project building with Java & Maven.
- Visualization with Grafana.
- Docker employed for project containerization.

Technologies used :

- Java, Spark, Kafka, PostgreSQL, Grafana, Docker.

Batch Processing: ETL pipeline, data modeling and warehousing of Sales data

https://github.com/anhvuphamtan/Batch-Processing

Objective :

- Utilize data collected from an e-commerce company's 2022 sales to **analyze their business performance.**
- **Design data models** for relational database and data warehouse (star schema).
- Develop an ETL pipeline to **transform raw data into actionable insights** then load to OLTP database, also store them in staging area.
- Implement a secondary ETL pipeline which **transform data from staging area into data warehouse** for enhanced data analytics. Visualize result.

Key features :

- ETL pipelines built with Python.
- Utilize PostgreSQL as OLTP database.
- Utilize S3 as staging area and Redshift as the data warehouse.
- Airflow for orchestration of pipeline workflow.
- Terraform for AWS Redshift provisioning.
- Docker employed for project containerization.

Technologies used :

- Python, PostgreSQL, Airflow, Terraform, S3, Redshift, Docker.