LUONG NGUYEN YEN NHI

Data Engineer

Ho Chi Minh City • (+84) 334 110 133 • lnyn.netlify.app • ynnhi1508@gmail.com

SUMMARY

I am a passionate Data Engineer who enjoys transforming raw data into scalable solutions. With experience building data pipelines, and working with Python, SQL, Spark, Kafka, Airflow, as well as data warehouses and data lakes, I'm driven by a constant desire to optimize performance and deliver real-world impact. I thrive in environments that challenge me to learn new technologies quickly and apply them effectively.

TECHNICAL SKILLS

- **Programming:** Python
- **Database:** PostgreSQL, MySQL
- **Big Data Tools:** Spark, Kafka
- **Orchestration:** Airflow
- **Cloud:** Azure (DP-203 Knowledge)

PROJECTS

Coffee Chain Data Platform | Python, MySQL, Kafka, Redis, Spark, Airflow, Minio, Prometheus, Grafana

June 2025 *Source: Project Website*

Goal: Built a hybrid pipeline for real-time promotion recommendations and batch analytics, enhancing customer engagement and business decision-making for a coffee chain.

- Processed ~1,500 orders per minute with ~5ms response time for real-time promotion.
- Used Kafka Connect to enable real-time CDC from MySQL for streaming new orders.
- Orchestrated daily batch jobs with Airflow and Spark, applying Lakehouse Medallion Architecture and Slowly Changing Dimension Type 2 for historical tracking.
- Automated data quality checks with Deequ to ensure reliable insights for the coffee chain.
- Integrated monitoring and alerting using Prometheus and Grafana.

Automated Job Data Pipeline | Python, Playwright, Postgres, Airflow

October 2024 *Source: Project Website*

Goal: Automate job data scraping and streamline storage with daily updates for better job monitoring.

- Scraped job data from TOPCV using Requests and Playwright.
- Cleaned and stored the data in PostgreSQL, generating a "due date" field based on the application deadline.
- Automated updates for the "remaining time to apply" field daily (e.g., "24 days left" to "23 days left").
- Orchestrated the entire pipeline (scraping, cleaning, and storage) with Airflow for seamless automation.

Real-Time NYC Trip Analytics | Python, Postgres, Kafka, Spark, PowerBI, Docker

August 2024 *Source: Project Website*

Goal: Improved trip monitoring and anomaly detection, providing actionable insights for NYC transport services.

- Deployed Dockerized Kafka for data streaming and Spark for real-time transformation and analytics.
- Analyzed trip trends and detected anomalies (e.g., short trips, fare mismatches).
- Stored processed data in PostgreSQL and visualized insights via an interactive PowerBI dashboard.