Tran Minh Duc

☎ 0973399776 ✉ minhductranit@gmail.com LinkedIn Github

EDUCATION

Ho Chi Minh City University of Technology and Education - HCMUTE Graduate Oct 2024
Major: Data Engineering *Ho Chi Minh City, Vietnam*

UniGap Academy May - Nov 2024
Data Coaching - Big Data *Ho Chi Minh City, Vietnam*

EXPERIENCE

Software Engineer Oct 2023 - Dec 2023
FPT Software *Ho Chi Minh City, Vietnam*

- Received tasks from the mentor and reported daily progress following the Scrum/Agile process.
- Designed data models to optimize CRUD operations for users, courses, classes, and training programs in a MySQL database.
- Developed user login and authentication API using the Spring Boot framework and managed source code with GitLab.
- Packaged the application and MySQL database with Docker for streamlined cloud deployment.

Data Engineer Nov 2024 - Present 2025
Viemind Technical Consulting *Ho Chi Minh City, Vietnam*

- Designed and implemented ETL pipelines using Dagster to read PDF, DOCX, and image data from a MinIO data lake, extracting text from, chunking it, and loading into a Qdrant vector database to power AI Agent knowledge bases.
- Developed a FastAPI-based back-end knowledge service to enable AI Agent queries via dense, sparse, and hybrid search modes and integrating with Dagster pipelines for dynamic knowledge base updates.
- Optimized data processing pipelines, reducing processing time by 60% through efficient code refactoring, resource management, and text chunking strategies.
- Containerized all pipelines and services with Docker and deployed them on Azure Kubernetes Service (AKS). Set up GitHub Actions CI/CD pipelines to automate testing and deployment, which reduced deployment time by 40% and ensured consistent rollouts.
- Collaborated with the AI team to develop ReactAgent and Agentic RAG systems, contributing to improved AI response accuracy.

PROJECTS

Real-Time User Behavior Streaming | *Spark, Kafka, Postgres, PySpark, Airflow, Docker* **[Github]**

- Integrated user behavior data from Kafka into a Postgres data warehouse for view analysis.
- Designed the warehouse with a Star Schema to optimize data structure.
- Used Spark for parallel reads from Kafka and dimension aggregation to improve query efficiency.
- Wrote data in parallel to Postgres to ensure storage performance.
- Automated end-to-end pipelines with Airflow, including email alerts for pipeline failures.
- Deployed Spark, Kafka, Airflow and Postgres with Docker for easier management and deployment.

Ecommerce User Behavior Data Pipeline | *DBT, SQL, Cloud Storage, Cloud Function, BigQuery, Looker* **[Github]**

- Built a user behavior data warehouse on BigQuery for analysis and visualization.
- Exported 30GB of user data, with over 40 million events, from MongoDB to Google Cloud Storage as a data lake.
- Used Cloud Function to automatically load data into BigQuery staging layer when added to the datalake.
- Crawled product images and names with BeautifulSoup, then loaded them into the datalake and BigQuery staging.
- Transformed staging data to Data Warehouse with DBT and visualized online sales in Looker.

Building a Data Warehouse for Online Retail | *Python, Pandas, DBT, Airflow, BigQuery, Docker* **[Github]**

- Loaded the Online Retail dataset from a CSV into BigQuery for efficient analysis.
- Built a data warehouse with DBT using a star schema for reporting.
- Automated the pipeline with Airflow for scheduled data processing.
- Containerized the technologies with Docker for easy deployment.

Chat with Your Document Using a RAG Engine | *Llama3, Chroma, Transformer, FastAPI, Streamlit* **[Github]**

- Built a chatbot that answers questions based on the documents users imported.
- Scan and extract text data from Word, PDF, and TXT files.
- Built a pipeline combining vector databases and large language models to generate answers based on the provided data.
- Built the back-end with FastAPI and the user interface with Streamlit.

SKILLS

Programming language: Python, SQL.

Database: MySQL, Postgres and MongoDB

Big data tools: Spark, Kafka, Databrick

ETL Tools: Airflow, Dagster

Basic knowledge of Google Cloud Platform, MS Azure

Visualization: Power BI

DevOps: Git, Docker, Kubernetes, CI/CD

English (Intermediate)

Certificates

- SQL (Advanced) Certificate - Hackerank `[View Certificate]`
- Python for Data Science, AI and Development `[View Certificate]`
- Python Project for Data Engineering `[View Certificate]`
- TOEIC certificate - IIG Vietnam