Nguyen Thanh Dat

Ho Chi Minh City

(+84) 859.229.215
nguyendat19222002@gmail.com
linkedin.com/in/nguyendat19
github.com/kuromine19

ABOUT ME

An enthusiastic and life-long learner seeking opportunities to gain hands-on experience related to Data Analysis/Data Engineering. Possess a strong foundation in data analytics and a passion for deriving meaningful insights from complex datasets. In the next three years, my objective is to become a professional Data Engineer.

EDUCATION

2020 - 2024

University of Science - VNU HCM

Computer Science - GPA: 8.27/10

EXPERIENCE

December 2024 - Present

Asia Commercial Bank – Analytics Engineer

Tools: SSMS (SQL Server), PLSQL (Oracle), Tableau, Metabase, Jira, Airflow

Languages: Python, PySpark, SQL

Key Responsibilities:

- **Data Pipeline Development:** Designed, built, and maintained ETL/ELT pipelines to ingest, process, and store data from diverse sources (raw files, APIs, streaming feeds).
- **Database Optimization**: Administered SQL Server and Oracle databases, optimized stored procedures and queries, implemented partitioning, and ensured high-performance data operations.
- **Workflow Orchestration:** Automated and monitored data workflows using Airflow, ensuring reliability and scalability of recurring jobs.
- **Data Delivery:** Provided clean, structured datasets to support business projects, reporting systems, and analytical teams.
- **Big Data Processing:** Leveraged PySpark and Python for distributed data processing, enabling efficient handling of large-scale datasets.
- **Monitoring & Troubleshooting:** Tracked pipeline performance, resolved system issues, and enhanced reliability of data infrastructure.

April 2024 - November 2024

Colgate - Palmolive Vietnam – Analytics Engineer Intern

Tools: Snowflake, Nifi, Domo, Google Apps Script, Confluence

Languages: Python, SQL, PySpark

Key Responsibilities:

- **Streamlining Processes**: Enhanced operational efficiency by automating repetitive business tasks.
- **Production Optimization**: Analyzed workflows to identify bottlenecks and minimize waste, boosting overall production effectiveness.
- **Quality Assurance Improvement**: Leveraged data analytics to uncover patterns in defects and rework, ensuring higher product quality.
- **Predictive Maintenance Support**: Contributed to initiatives aimed at reducing equipment downtime and associated costs through predictive analytics.
- **Dashboard Development**: Created interactive dashboards to visualize key metrics and continuously refined them to align with evolving business goals.
- **Data Governance Assistance**: Supported internal audit processes to ensure robust data governance practices.

SKILL

Technical Skills

Data Acquisition	Selenium, RESTful APIs, Data Link
Processing Languages	Python, SQL
Relational Database	MSSQL, MySQL, Oracle
NoSQL Database	MongoDB, DynamoDB
Data Warehousing	RedShift, Snowflake
BI Tools	Tableau, Domo, Metabase
Big Data & ETL Technologies	Hadoop, PySpark, Kafka, Airflow, Nifi, dbt
Cloud Computing	AWS (EC2, S3, RDS, Glue, Lambda)
Version Control	Git
Other Skills	HTML, CSS, JavaScript, Django, Git, Google Apps Script

Soft Skills

English (TOEIC: 870/990)
Problem-solving
Analytical & Critical thinking
Adaptability in a fast-paced environment
Teamwork, Communication & Presentation

PROJECTS

Employee Attrition Prediction Pipeline
February 2026 – June 2026

Description: Developed a scalable data pipeline to prepare HR datasets for machine learning models predicting employee attrition. Automated feature engineering, data cleaning, and model retraining workflows to ensure consistent and reliable outputs.

- **Tools:** Airflow, Metabase
- **Languages:** Python, SQL
- **Frameworks:** PySpark
- **DBMS:** Oracle

Role:

- Designed and implemented ETL pipelines using Airflow to automate ingestion and transformation
- Applied PySpark for distributed data processing, enabling efficient handling of large-scale employee records.
- Automated model retraining and deployment, ensuring predictions remained up-to-date with new data.
- Integrated curated datasets into Metabase for business teams to monitor attrition risk.

Achievements:

- Reduced manual preparation time for HR analytics by 80% through automated pipelines.
- Cut retraining cycle time from 1 week to 1 day, enabling faster adaptation to workforce changes.
- Enhanced business decision-making by providing daily attrition risk dashboards to HR managers.

HR Data Optimization
March 2025 – July 2025

Description: Standardized HR data retrieval principles to ensure data quality and optimize query performance.

- **Tools:** Metabase, Airflow
- **Languages:** SQL, Python
- **DBMS:** SQL Server

Role:

- Reviewed and standardized SQL queries, ensuring accuracy, reliability, and consistency of HR data retrieval.
- Optimized complex queries and restructured formulas to significantly reduce

execution time.
- Built automated ETL workflows to prepare curated datasets for HR reporting and analytics.
- Developed secure data access solutions in Metabase, preventing direct exposure of raw data sources while enabling flexible reporting.
- Implemented validation and normalization rules to maintain data integrity and compliance with audit policies.

Achievements:
- Reduced tenure calculation query time from 1 hour 30 minutes to just 2-3 seconds.
- Created an up-to-date employee list with query time reduced from over 5 minutes to 1-2 second.
- Increased efficiency and governance by enforcing controlled access through views/procedures, improving both security and usability.
- Enhanced HR reporting capabilities by delivering optimized datasets that supported faster and more reliable analytics.

Most-played Games Dashboard
Oct 2023 - Dec 2023

Source: *https://bom.so/06TsQy*

Demo: *https://bom.so/hRPz3G*

Description: Designed and implemented a scalable data pipeline to continuously ingest and process gameplay statistics from Steam, enabling monitoring of most-played games. Built a robust infrastructure to support downstream analytics and dashboards.

- **Tools**: Selenium, GitHub Actions, Chart.js
- **Languages**: Python, SQL, JavaScript
- **Frameworks**: Django, Bootstrap
- **DBMS**: MongoDB Atlas

Role:
- Developed automated data crawling pipelines using Selenium and GitHub Actions to collect player activity data from Steam.
- Designed and managed MongoDB Atlas schema to store raw and processed data efficiently.
- Implemented ETL workflows with Python and Airflow to clean, transform, and aggregate gameplay statistics.
- Built APIs in Django to serve processed datasets to downstream applications.
- Integrated Chart.js to visualize real-time metrics in interactive dashboards for game operators.

Achievements:
- Established a fully automated pipeline capable of handling continuous data ingestion without manual intervention.

Sentiment Analysis
February 2023 - April 2023

Source: *https://bom.so/dtdRRI*

Demo: *https://bom.so/Pslp1b*

Description: Built a scalable data ingestion and processing pipeline to support a sentiment analysis model predicting emotions expressed in tweets. Focused on automating data collection, preprocessing, and deployment infrastructure to ensure reliable and efficient model integration.

- **Tools**: Hugging Face
- **Languages**: Python, SQL
- **Frameworks**: Streamlit
- **DBMS**: PostgreSQL

Role:
- Designed and implemented an ETL pipeline to ingest tweets via APIs, clean and normalize text data, and store results in PostgreSQL.
- Integrated Hugging Face BERT model into the pipeline, ensuring seamless deployment and updates.

- Built a Streamlit web app connected to the pipeline, enabling real-time sentiment predictions for end-users.

Achievements:

- Automated retraining cycles, cutting manual intervention by 90% and ensuring models stayed current with new tweet data.
- Delivered a production-ready sentiment analysis dashboard that supported predictions for thousands of tweets daily.