Hoai Nam Nguyen Cong

Thu Duc, Ho Chi Minh City | nchn.work@gmail.com | (+84) 838 380 386 | Website | LinkedIn | GitHub

Introduction

Data & AI enthusiast with hands-on experience in building data pipelines and AI agents. Passionate about turning data into actionable insights that drive real-world impact.

Experience

VietData.AI, Fresher AI Engineer — March 2025 - Now

- Built a real-time pipeline to extract structured data from PDF CVs using **Microsoft Graph API** and optimized **OpenAI prompts**; integrated with AI search for instant recruiter access.
- Developed a **RAG-based AI chatbot** for talent acquisition with **LangChain**, **OpenAI API**, and **Hugging Face**, boosting search accuracy and recruiter productivity.
- Implemented **CI/CD with GitHub Actions**, managed **Cloudflare** and **Nginx** for secure deployments across production and staging.
- Leveraged **Google Cloud Platform** for scalable data processing, storage, and analytics in diverse data projects.
- Worked closely with stakeholders to deliver tailored AI solutions aligned with business needs.

AISIA Research Lab, Research Assistant — January 2025 - Now

- Built pipelines for collecting, preprocessing, and annotating **100+ hours of Vietnamese speech data** for **Text-to-Speech** and **Speech-to-Text** model training.
- Applied **VAD** to remove non-speech, enhancing data quality.
- Used **Speaker Diarization** with **90%+ accuracy** to segment multi-speaker recordings.
- Optimized segmentation via **clustering**, reducing labeling effort by **20%**.
- Delivered higher model performance through cleaner, well-segmented data.

Education

VNU-HCM, University of Science, Bachelor of Data Science — August 2021 - June 2025 (Expected)

- **GPA**: **3.3/4.0**
- **Coursework:** Object-Oriented Programming, Data Structures and Algorithms, Database Management Systems, Machine Learning, Deep Learning, Big Data, Data Visualization, Artificial Intelligence, Data Mining.

Language

- **TOEIC Listening and Reading: 900/990**
- **VSTEP (Vietnamese Standardized Test of English Proficiency): Level 4/6 (B2)**

Technical Skills

- **Big Data Technologies:** Apache Spark, Apache Hadoop, DBT, Dagster
- **Storage Solutions:** MinIO (S3-compatible), PostgreSQL, MySQL, MongoDB, BigQuery, CloudSQL, GCS
- **Machine Learning:** Scikit-learn, Tensorflow, Keras, Pytorch
- **Computer Vision:** OpenCV, YOLO, DeepSORT
- **Cloud Platforms:** Google Cloud Platform, Microsoft Azure (Graph API, Portal)
- **DevOps and Deployment:** GitHub Actions, Docker, Nginx, Cloudflare, CI/CD, SSH, Unix/Linux, REST API, WebSocket
- **Programming Languages:** Python, SQL, R, C/C++
- **Visualization and Monitoring:** Power BI, Metabase, Grafana, Streamlit, Prometheus

Projects

Tiki Recommender ETL Pipeline

- **Abstract:** End-to-end ETL data pipeline that automatically scrapes data from **Tiki.vn**, stores it in **MinIO (data lake)**, transforms data and trains the model using **Apache Spark**, and loads the processed data into **PostgreSQL (data warehouse)**, orchestrated by **Dagster** and containerized with **Docker**.
- **Serving:** Designed a dashboard using **Metabase (BI tool)** to extract valuable insights and deployed a **Streamlit app** to provide personalized product recommendations for e-commerce users.
- **Website:** tiki-recommender-etl-pipeline.streamlit.app
- **Source Code:** github.com/nchn471/tiki-recommender-etl-pipeline

Comprehensive Performance Analysis of Hadoop and Spark

- **Abstract:** This study developed a fully distributed system on Linux virtual machines using Hadoop and Spark, benchmarking various parameters to optimize performance and evaluate the efficiency of **Apache Hadoop** and **Apache Spark** in large-scale data processing.
- **Workloads:** WordCount, TeraSort.
- **Tools Used:** Hadoop, Spark, HiBench, Prometheus, Grafana, Unix/Linux.
- **Vietnamese Report:** Check it here
- **Source Code:** github.com/nchn471/performance-analysis-hadoop-spark

Developer Salary Prediction Website

- **Abstract:** A machine learning project that predicts developer salaries based on experience, country, education, and technology stack. The project also includes visualizations to analyze developer's market trends using the Stack Overflow 2023 Survey Dataset.
- **Tools Used:** Python, Streamlit, Matplotlib, Seaborn, Scikit-learn, NumPy, Pandas, Joblib.
- **Website:** developer-salary-prediction-project.streamlit.app
- **Source Code:** github.com/nchn471/Developer-Salary-Prediction

Certificates

- **Professional Machine Learning Engineer** by **Google Cloud Platform**
- **Fundamentals Data Engineering** by **AIDE Institute**