(+84) 382034536

Hanoi, Vietnam

anhtuanhoang2001@gmail.com

github.com/tuanx18/data-engineer-portfolio

linkedin.com/in/tuanx18

Anh Tuan Hoang

A recent graduate holding a Bachelor's degree in **Software Application and Technology,** my skill set spans *data manipulation, big data handling,* and *programming*. Proficient in *SQL, PowerBI, Python, Apache Spark, Hadoop ecosystem,* and *MongoDB* allows me to lead ETL processes, design robust Data Pipelines, and optimize Data Warehouses for scalable and efficient data management. Equipped a keen programming attitude developed during my university time learning about Software and a natural tendency for leadership, desired to contribute technical expertise and a collaborative spirit to drive impactful data-driven initiatives in a dynamic **Data Engineering** role.

EDUCATION

Bachelor of Information and Communication Technology (Software Technology)   *Jan 2020 – Nov 2023*
SWINBURNE UNIVERSITY OF TECHNOLOGY
Courses: Database Analysis and Design, User-Centred Design, Introduction to Data Science, Introduction to Artificial Intelligence, Big Data Management, Cloud Computing Architecture, Professional Issues in Information Technology

Computer Science   *Aug 2019 – Jan 2020*
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

TECHNICAL EXPERIENCE

Data Engineer, Scrum Member   *May 2023 – Aug 2023*
Novaland Data Gathering and Consolidated Report Automation Project
- Orchestrated data extraction procedures involving a chatbot leveraging NLP, a Python-based program for mining insights from phone call data, and a Google Form for user manual inputs.
- Crafted the architecture for the Data Pipeline, establishing a local NoSQL database system as the central Data Lake. This hub streamlined the consolidation of data from diverse sources (chatbot, phone calls extraction and forms)
- Developed a prototype for ETL (Extract, Transform, Load) data processing, automating the refinement and organization of data within the Novaland's Data Warehouse, preparing it for utilization by the organization.

Data Specialist, Intern   *Aug 2022 – Dec 2022*
FPT Software
- Worked as a team of seven, built samples and useful databases data specialists learned and researched database management and data insights.
- Mainly working on SQL, Python, Oracle, AWS and Microsoft Database Access

SKILLS AND CERTIFICATES

Programming Languages:	**Python**, Kotlin, C++, JavaScript, HTML / CSS, SQL
Technologies	Pandas, Data Visualization (PowerBI, Tableau), SQL Databases (SQL Server, AWS RDS, MySQL, PostgreSQL), NoSQL Databases (MongoDB, Redis), Batch Processing (Spark, Hadoop, Hive), Apache Airflow, ETL Data Pipeline Processes, Amazon Web Services, Git
Others	Machine Learning Models, AI path-finding Algorithms, Object-Oriented Programming, Data Structures & Patterns, Agile Scrum Methodology, **IELTS Overall Band 7.5**

PROJECTS   [Click to access my Full Portfolio ✓]

Billionaire Research and Exploration   *Oct 2023*
- Conducted comprehensive analysis leveraging Scala Spark and PySpark and the Billionaires Dataset to explore the realm of extreme wealth.
- Utilized PySpark's Machine Learning Models to implement machine learning algorithms, including K-means clustering and logistic regression.

Technologies: Python, Pandas, Matplotlib, Apache Spark, Confusion Matrix, Machine Learning Classifier Models (MLlib)

IT Department Project, Swinburne University *May 2022 – Aug 2022*
Database Developer
- Organized and Set up Database for the Student Council within the university using Microsoft Database Access
- Managed Lower-level Data Specialists to ensure the dataflow run smoothly

Technologies: Microsoft Access, SQL, Python, Pandas, Machine Learning

Stellar Classification *Mar 2021*
- Equipped with a dataset containing 18 curated feature columns, our project holds the promise of advancing astronomical science and discovery in several remarkable ways.
- Aimed to uncover the mysteries of the universe by classifying celestial objects such as stars, quasars, and galaxies.

Technologies: Python, Pandas, Matplotlib, Correlation Heatmap, Confusion Matrix, Classification Report, ROC Curve, Machine Learning Classifier Models.