Mai Dang Nhat Anh

0971538165 | mdna.tl2112@gmail.com | github.com/mdnanh

As a student passionate about data, I am always eager to learn and grow in this field. Beyond the foundational knowledge gained in school, I actively self-study and explore new technologies. I am seeking a professional environment where I can apply my skills while learning new concepts to solve real-world business data challenges and contribute to the company's project.

Education

International University — Vietnam National University, HCMC

Bachelor of Data Science — *Class of 2019*

Coursework: Data Analysis, Fundamentals of Big Data Technology, Data Science and Data Visualization, Principles of Database Management, Algorithms & Data Structures.

Projects

Real-time E-commerce Data Streaming | github.com/mdnanh/StreamingEcommerceDE

|*PySpark, Kafka, Druid, MiniO, Airflow, Superset, Docker*|

- Designed and implemented a real-time data streaming and analysis system for an e-commerce platform using Apache Kafka, Druid, and MinIO, enabling efficient processing of large-scale, multi-source data.
- Developed dashboards with Apache Superset and orchestrated complex data workflows using Apache Airflow
- Improved decision-making by providing real-time insights via interactive dashboards on Apache Superset.
- Engineered a scalable architecture using Docker containerization, ensuring seamless deployment, efficient resource utilization.
- Data collected from Tiki e-commerce platform.

User Behavior Data Pipeline for Smart TV Analytics | github.com/mdnanh/CustomerAnalysisBigData

|*PySpark, MySQL, Json, PowerBI*|

- Designed and implemented a data pipeline for analyzing user behavior on Smart TV platforms, processing 30 JSON files (8GB) containing viewing and search activity data.
- Utilized Apache Spark to handle large-scale data processing, including cleaning, transformation, and aggregation, ensuring data consistency and quality.
- Developed trend analysis algorithms to track changes in customer content preferences over time, enabling businesses to understand shifting user behaviors.
- Built and deployed interactive dashboards in Power BI, visualizing monthly trends in popular content categories and key user engagement metrics.
- Delivered actionable insights through automated reports, helping stakeholders make data-driven decisions to optimize content strategy.

CDC Data Pipeline for Recruitment Platform | github.com/mdnanh/CDCAnalysisRecruitment

|*PySpark, Kafka, Cassandra, MySQL, Local Storage, Parquet, Json, Docker*|

- Designed and developed a Change Data Capture (CDC) data pipeline to support analytical and operational processes for a recruitment platform.
- Processed and transformed 28 folders of Parquet files (150MB) stored locally, ensuring efficient data ingestion and incremental updates.
- Built and optimized ETL workflows, including data cleaning, duplication, and schema validation, to ensure data accuracy and consistency in the pipeline.
- Integrated processed data into a MySQL database, enabling easy querying and supporting downstream analytics.
- Optimized data processing workflows to work efficiently on limited hardware resources, improving overall processing efficiency and ensuring smoother operation on weaker systems.

Experience

AI Engineer Intern Oct. 2023 – Jan. 2024

Bitdance Technologies and Entertainment Company *Ho Chi Minh city*

- Researched the application of large language models (LLMs) in content analysis and technical aspects of music creation, exploring innovative techniques for leveraging AI in the music domain.
- Researched sentence embedding techniques and developed retrieval-based question answering systems to implement Retrieval-Augmented Generation (RAG) pipelines.
- Contributed to optimizing and evaluating embedding models to enhance the accuracy and efficiency of the retrieval system.

Data Science Intern May. 2023 – Aug. 2023

Stech Technology Joint Stock Company *Ho Chi Minh city*

- Crawled product review data from e-commerce platforms to create a labeled dataset for training and evaluating models.
- Conducted research and developed sentiment classification models and aspect-based sentiment analysis to identify key factors influencing positive or negative evaluations in textual data.
- Acquired and applied self-learning, research methodologies, and technical report presentation skills in the domain of NLP.

Technical Skills

Programming Languages: Python, SQL
Data Engineer: ETL, ELT, Data Modeling, Apache Airflow
Big Data Tools: Apache Spark, Kafka, Druid
Database: MySQL, PostgresSQL, Cassandra, MongoDB
Data Lake: MiniO
Visualize Tools: Apache Superset, Grafana, PowerBI
Other Tools: Git, Docker