Duy Ha Van

2225 Pham The Hien, Ward 6, District 8 • Ho Chi Minh City, Vietnam
hvduy37@gmail.com • 0396161438 • https://github.com/viplazylmht

Education

UNIVERSITY OF SCIENCE, VNU-HCM Ho Chi Minh City, Vietnam
Bachelor, Data Science in Computer Science. GPA **8.5** May - 2022

Skills & Interests

Technical:

- **Programming Languages**: Python, SQL, C++, R, Java
- **ETL Platforms**: Apache Airflow, dbt
- **Bigdata Platform**: Apache Spark, Trino, Bigquery, Delta Lake, Apache Iceberg
- **Cloud Service**: Google Cloud (Bigquery, GCS, Dataproc, GKE, Cloud Function, Identity), Oracle (Oracle APEX)
- **Operating Systems**: Linux, Windows, MacOS
- **Data Ops**: CI/CD, Github, Gitlab, Docker, Kubernetes
- **Data Governance**: Apache Ranger, Great Expectations, Datahub, SOC 2-compliant
- **Artificial Intelligence**: Machine learning, Generative AI, Agentic AI, LangChain
- **Soft Skills**: Time management, Self-learning, Communication, Teamwork
- **Others**: Cyber security

Soft skills:

- Time management: organization, planning, goal setting
- Adaptability: self-management, self-motivation
- Problem-solving skills: observation, analysis, brainstorming, and decision making.
- Creativity, positive energy and attention to detail.
- Teamwork
- Communication: presentation, discussion, and active listening

Language: English (Low Intermediate)
Interests: Photography, music, running, and open sources

Experience

MOMO Ho Chi Minh City, Vietnam
Data Engineer - MoMo Talents Program Jan 2022 – Present

- **Responsibilities**
  - Built ETL pipelines to ingest data from various sources (Oracle, Big Query, GCS data lake, …) to Big Query data warehouse.
  - Built data models and ETL pipelines for business reports based on user's requirements.
  - Improved data quality by developing a tool and service to help other departments monitor and receive automatic alerts for data quality issues.
  - Managed Bigquery resource allocation across the entire platform.
  - Optimized queries and services to save 40% of cost without any stuck workload.
  - Designed and implemented a lakehouse solution to reduce cost of all workloads.
  - Developed an end-to-end transpiling tool to translate queries between the data warehouse and lakehouse.
  - Reduced human labor costs by up to 90% in the pipeline migration process using the transpiling tool.
- **Technologies**
  - Workflow Orchestration: Apache Airflow
  - Container Orchestration: Kubernetes

  - Build tool and Containerization: Bazel, Docker
  - CI/CD: Github Actions, Gitlab Workflow, Jenkins (fundamental)
  - Cloud Service: Google Bigquery, GCS, Google Dataproc, GKE, Google Cloud Function, Google Identity Platform, Oracle APEX, …
  - Big data processing: Bigquery, Apache Spark, Trino (distributed SQL query engine)
  - Transformation tool: dbt
  - Security: Oauth2, OpenID Connect
  - Programming Language: Python, SQL, Java
  - Artificial Intelligence: Generative AI, Agentic AI, LangChain

Project

MOMO Data Engineer

Data Agent

- **Project description**
  - Developing GenAI and Agentic AI agents to help users quickly extract insights from internal data and documents.
- **Responsibilities**
  - Design & Implementation: Build a scalable, maintainable, and extensible codebase to support engineers in developing new AI agents.
  - Research & Experimentation: Explore autonomous decision-making capabilities in agentic AI designs.
  - Model Development: Develop foundation models based on agentic AI principles for various business applications.
  - Practical Optimization: Fine-tune AI systems to align with business objectives while ensuring ethical and responsible AI practices.
  - Collaboration: Work with cross-functional teams to align AI solutions with business needs and enable seamless deployment.
- **Goal**:
  - Reduce engineers' time spent on periodic data analysis by 80%.
  - Enable autonomous AI-generated insights for customer reports.
  - Develop a chatbot for engineers and customers to easily query and extract insights about their data and documents.
- **Technologies**
  - GenAI, Agentic AI, LangChain, SMTP Email, FastAPI, Chatbots

MOMO Data Engineer

Access Management – Data Security

- **Project description**
  - Develop a SOC 2-compliant platform to manage time-based privileged access to sensitive data and policy tags across multiple data warehouses, data lakehouses, and services.
- **Responsibilities**
  - Design and implement time-based privileged access control and policy tagging across data warehouses, lakehouses, and services.
  - Ensure SOC 2 compliance by enforcing access policies, auditing, and generating compliance reports.
  - Collaborate with security and data teams to align governance, access rules, and monitoring requirements.
  - Build and maintain logging, monitoring, and automation for access approval, revocation, and expiration processes.
- **Goal**: The Access Management Tool centralizes the approval process for 100% data access requests within the data platform.
- **Technologies**

    - SOC 2-compliant, SMTP Email, FastAPI, OpenID Connect

MOMO Data Engineer

Data Pipeline Migration

- **Project description**
    - Migrate data platform and departmental workloads to the new data lakehouse.
- **Responsibilities**
    - Developed a transpiling tool using open-source projects to facilitate the migration of SQL from the current production environment to the Lakehouse.
- **Goal**: The transpiling tool reduced migration costs by up to 90% at Momo.
- **Technologies**
    - SQLGlot, Trino/Presto, Bigquery, Airflow, OpenID Connect

MOMO Data Engineer

Data Lakehouse – Data Ops

- **Project description**
    - Research and implement a cutting-edge data solution to stay current with industry trends, centralize workloads, and reduce BigQuery costs.
- **Responsibilities**
    - Designed and implemented a lakehouse solution to reduce cost of all workloads.
    - Evaluated various open file formats and selected the most suitable one for use in the lakehouse.
    - Integrated the Lakehouse with new access management systems to enhance data security.
    - Migrated core ETL pipelines to the lakehouse.
- **Goal**:
    - Trino runs on GKE as a distributed query engine to process large batch data stored in GCS.
    - Reduce up to **70% cost** per workload thanks to spot instance without any data SLA.
- **Technologies**
    - Trino, Spark, Apache Ranger, GKE, GCS, Bigquery Storage, dbt, and Airflow

MOMO Data Engineer

Data Observability – Data Governance

- **Project description**
    - Reduce the workload of the data-platform team in responsiveness to data for both info and incident.
- **Responsibilities**
    - Implemented a system based on popular open-source projects which helps end-user monitor five pillars of data: Freshness, Volume, Quality, Schema, and Lineage.
- **Technologies**
    - Datahub, dbt, Great Expectations, and Airflow

MOMO Data Engineer

Cost Optimization – Reduce cost on GCP

- **Project description**
    - Reduce GCP costs as much as possible in response to economic downturns and changes in GCP billing policies.
- **Responsibilities**
    - Supported other departments in optimizing queries.
    - Moved services, ETL, and ELT pipelines to on-premises Kubernetes.
    - Experimented shifting from Google Bigquery to Vertica.
    - Managed GCP resources for each team and department by the divide-and-conquer principle.

- **Goal**: Saved **40%** GCP cost without any stuck workload.
- **Technologies**
    - Bigquery, Vertica, Kubernetes, Oracle APEX, and GCP gRPC API

MOMO Data Engineer

Golden Record - Process for developing a high-value Data Mart

- **Project description**
    - Develop a streamlined process to assist other departments in creating high-value reports for both internal teams and merchants.
- **Responsibilities**
    - Researched and built a data quality tool on top of open-source Great Expectations project to control the data model's quality, freshness, and extensionality.
    - Guided other departments in all steps of golden record project, especially the ensure data quality step.
- **Goal**: Served many dataflows such as events and transactions of the MoMo Super App.
- **Technologies**
    - dbt, Great Expectations, Airflow, Gitlab, Kubernetes, Oracle OCI, and Oracle APEX

Publication

MEP: A Comprehensive Medicines Extraction System on Prescriptions ICCCI 2023

Conference paper | First Online: 13 September 2023 Computational Collective Intelligence
pp 713–725

Medical Prescription Recognition Using Heuristic Clustering and Similarity Search ICCCI 2022

Conference paper | First Online: 21 September 2022 Computational Collective Intelligence
pp 768–780

Contribution

SQLGlot - Contributing to the open-source SQL parser

- **Project description**
    - SQLGlot is a no-dependency SQL parser, transpiler, optimizer, and engine. It can be used to format SQL or translate between 21 different dialects.
- **Responsibilities**
    - Improved the accuracy of translation between Bigquery and other SQL dialects.
- **Technologies**
    - SQLGlot, Trino/Presto, Bigquery

Great Expectations – Contributing to the open-source data quality project

- **Project description**
    - GX is an open-sources project to validate and monitor the quality and freshness of data.
- **Responsibilities**
    - Supported the new Vertica dialect, enabling Great Expectations to assess data quality on the Vertica database.
- **Technologies**
    - Great Expectations, Vertica, Bigquery