HÀ VĂN DUY

Data Engineer

SKILLS

- Team Work / Creative
- Scrum / Agile
- Docker / Kubernetes
- CI / CD / Shell / Linux
- ETL / ELT / Data modeling
- Cloud (GCP)
- Data Governance
- Generative AI

LANGS / TOOLS

- C/C++, Java, Kotlin, Python
- Advanced SQL
- Great Expectations / Datahub
- dbt / Airflow / Oracle APEX
- Trino / Apache Ranger
- Delta Lake / Apache Iceberg

CONTACTS

hvduy37@gmail.com

+84 39 616 1438

Ho Chi Minh City

github.com/viplazylmht

INTRODUTION

I am currently a Data Engineer at Momo (M_service). Being a person of harmony, I am totally wanna face new challenges and take risks. My career gone is to succeed in the field of Big Data & AI. I am on my journey to acquire knowledge, down-to-earth experience, to gain the result I was always looking for.

EDUCATION

2018 - 05/2022 University of Science, VNU-HCM

Data Science Major - GPA: 8.5

EXPERIENCE

01/2022 - present
2yr 5mos

MoMo (M_Service)

Data Engineer - From MoMo Talents Program
Big Data & AI department, Data Platform team

PROJECTS

Golden Record - Process to achieve high-value Data Mart

Build tools and services on top of open-source projects to control the data model's quality, freshness, and extensionality. Golden Record currently serves many dataflows such as events and transactions of the MoMo Super App.
Used: dbt, Great Expectations, Airflow, Gitlab, Kubernetes, Oracle OCI, and Oracle APEX

Cost Optimization - Reduce cost on GCP

Collaborate with team to support other teams to optimize queries. Move services, ETL, and ELT to on-premise Kubernetes. Try to shift from Bigquery to Vertica. Manage GCP resources for each team in MoMo by the divide-and-conquer principle.
Conclusion: 40% cost saved without any stuck workload.
Fluent in: Bigquery, Vertica, Kubernetes, Oracle APEX, and GCP gRPC API

Data Observability - Data Governance

Just a project which helps end-user monitor five pillars of data: Freshness, Volume, Quality, Schema, and Lineage. This project aims to reduce the workload of the data-platform team in responsiveness to data for both info and incident.
Fluent in: Datahub, dbt, Great Expectations, and Airflow

Data Lakehouse

Collaborate with the team to build a lakehouse solution to reduce cost of all workloads at Momo. Trino/Spark run on GKE as a 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.
Fluent in: Trino, Spark, Apache Ranger, GKE, GCS, Bigquery Storage, dbt, and Airflow

Data Pipeline Migration

Build a transpiling tool based on top of open-source projects to help end-to-end migrate SQL from current production environment to the Lakehouse, reduce up to 90% human cost of the migration phase at Momo.
Fluent in: SQLGlot, Trino/Presto, Bigquery, Airflow

CONTRIBUTIONS

SQLGlot Contributing to the SQL transpiler

SQLGlot is a no-dependency SQL parser, transpiler, optimizer, and engine. It can be used to format SQL or translate between 21 different dialects.

Great Expectations Add support for Vertica dialect

GX is an open-sources project to validate and monitor the quality and freshness of data.

And so on which can be found in my github profile (QR code)

INTERESTS

PHOTOGRAPHY

MUSIC

ESPORT