Thoai Trinh Van

HCMC, Vietnam • trinhvanthoai99@gmail.com • +84 338 892 212 • linkedin.com/in/chimeyrock999
github.com/chimeyrock999 • chimeyrock.gitbook.io

Data Platform Engineer with hands-on experience building data products across metadata, governance, discoverability, and platform operations. Focused on making enterprise data platforms reliable, cost-efficient, and easy to use.

WORK EXPERIENCE

Data Platform Engineer **Mar 2026 - Present**
Masan Group **Ho Chi Minh, Vietnam**

- Built an AI-powered data discovery chatbot enabling users to explore available datasets by business domain, data layers, and end-to-end flows, with natural-language access to dataset inventory, ownership, and high-level table descriptions.
- Designed and implemented an LLM-powered data dictionary and metadata governance workflow on Databricks leveraging Unity Catalog for 10K+ datasets, generating draft table/column descriptions, business grain, ownership and governance tags, key field suggestions, and PII classifications with human-in-the-loop validation to improve catalog completeness and enable downstream data quality, masking, access control, and discovery use cases.
- Implemented metadata-driven observability platform to detect low-usage datasets, partition misconfiguration, and storage anomalies, saving ~$100/day in ADLS cost.

Data Engineer **Sep 2025 - Feb 2026**
Zalopay - VNG Corporation **Ho Chi Minh, Vietnam**

- **Designed** and **integrated** a unified **feature store** based on Feast into a legacy data platform, supporting batch and real-time feature engineering and API-based online serving, with GitOps-style governance and versioning; operated at scale with **14M MAUs**, **~2M** streaming **events/day**, and **<200 ms** feature retrieval **latency**.
- **Built** batch and **streaming ETL / feature engineering pipelines** using Spark and Airflow; **developed internal frameworks** and **libraries** to define & automate streaming pipeline deployment and management, allowing ML Engineers and Analysts to focus on business logic instead of infrastructure complexity.
- **Designed a declarative Spark-based framework for graph pipelines** , providing ORM-like abstractions for defining vertices, edges, and transformations with built-in validation and pluggable storage writers.
- **Built a large-scale User Network graph pipeline (~50M transactions/day)** to model relationships between users and support fraud investigation and risk analysis.
- **Owned** and **improved** the Risk data & ML platform (Spark, Airflow, HDFS, and related systems), **ensuring stable** daily operation of **50–60 Spark applications**, each processing **up to 500M–1B records** per run, through performance tuning, monitoring, and operational best practices.

Data Platform Engineer **Apr 2023 - Aug 2025**
FPT Smart Cloud - FPT Corporation **Ho Chi Minh, Vietnam**

- **Built** a **Lakehouse platform solution** (Iceberg + S3 + Spark + Trino) with **full governance**: OAuth2 propagation, Ranger FGAC + masking, OpenMetadata lineage, and S3 SSE-C encryption.
- **Implemented** CDC/Streaming pipelines (Kafka Connect + Debezium) ingesting **100GB/day & 5K TPS**, **synconizing 500+** PostgreSQL **tables** to ClickHouse, Iceberg & S3.
- Developed a self-service Spark environment via **JupyterHub** (**custom Profile Manager**), enabling **per-profile resource isolation**, **secure secret injection**, and **automatic Spark bootstrap**.
- Created **Airflow plugins** integrating **Spark Operator**, standardizing **job submission**, **dependency management**, and **real-time log streaming**.
- Built monitoring end-to-end solution (Prometheus + Grafana) tracking latency, throughput, error rate across Spark, Kafka, Airflow and others Data Platform services.
- Operated the **core platform stack** on **Kubernetes**: **Airflow**, **Trino**, **Superset**, and **Kafka Connect**, following a **GitOps style** with **ArgoCD**.

Backend Engineer Intern **Oct 2022 - Mar 2023**
FPT Smart Cloud - FPT Corporation **Hanoi, Viet Nam**

- Researched Kafka architecture and deployment feasibility, and designed Kafka-as-a-Service solutions on both VMs and Kubernetes.
- Deployed Kafka on Kubernetes using Strimzi and Implemented end-to-end monitoring with JMX, Telegraf, Prometheus, Grafana, and alerting through Telegram.
- Built custom enterprise-grade Kong plugins and integrated the API gateway into Kubernetes microservices, reducing licensing costs by replacing commercial plugin features with in-house implementations.

EDUCATION

Bachelor of Science (B.Sc.) in Computer Science **Sep 2018 - Sep 2023**
Hanoi University of Science and Technology

PROJECTS

JupyterHub Profile Manager – Microservice Extension for Access Control and Spark + Lakehouse Integration

- Built a production-grade **extension** on top of **open-source JupyterHub**, providing **ready-to-use Spark environments** that require **zero user configuration**.
- Delivered fully governed execution with **row-level access control** via Apache Ranger, **auto lineage** via OpenMetadata, và **native LakeHouse integration** (S3, Iceberg, metadata & auth binding).
- Implemented **OAuth2**, **RBAC**, **AES-256-GCM** secret encryption, and dynamic per-profile resource provisioning on **Kubernetes**.
- Automated Spark bootstrap so users can run Spark with correct credentials, policies, catalogs, and storage bindings **immediately on first launch**.
- Improved developer experience and onboarding speed; adopted internally and many enterprise clients (e.g FPT Retail, FPT IS, ATIS)

Airflow Plugin for Spark Job Management

- Developed a production-ready **Airflow plugin** enabling governed, **self-service Spark job submission** and scheduling directly from Airflow UI — **no Kubernetes or YAML exposure** for users.
- Provided a unified job submission UI supporting Python / Scala, multiple Spark versions, and automatic packaging via Pip, Maven, or pre-built environments.
- Ensured environment consistency by uploading dependencies to S3, eliminating drift across Spark clusters.
- Added **real-time job tracking** (state, logs, failures) and seamless navigation to **Spark UI / History Server** through a built-in **reverse proxy**.
- Improved developer experience by allowing users to submit, monitor, and re-run Spark jobs **within minutes**, reducing operational overhead and increasing adoption across teams.

ORGANIZATIONAL & VOLUNTEER EXPERIENCE

Open Source Contributor **Sep 2025 - Present**
feast.dev **Ho Chi Minh, Vietnam**

- Added **HDFS Staging support** for **Spark Offline Store**, enabling **distributed materialization** and more **efficient large-scale feature computation**.
- Introduced **HDFS Registry backend**, allowing teams to manage **Feast feature definitions** on **Hadoop-compatible file systems**.
- Optimized **MySQL Online Store write performance** by implementing **batch insert** and **transaction grouping**, significantly reducing **write latency**.

AWARDS

Best Performance Award – Q3 2024 by **FPT Smart Cloud** **Oct 2024**

Recognized as one of the top performers in Q3 2024 for outstanding contributions to Data Platform as a service projects.

SKILLS

Programming: Java, Python

Data Engineering: Spark, Airflow, Kafka, Debezium, Feast

Storage & DB: Iceberg, ClickHouse, TiDB, PostgreSQL, S3, HDFS

Cloud & DevOps: Kubernetes, Helm, Docker, GitLab CI/CD, ArgoCD

Monitoring: Prometheus, Grafana, Telegraf