Minh Pham

Email | LinkedIn | Portfolio | GitHub | 682-597-4780

Experience

Insurify **May 2025 – Present**

Analytics Engineer - Marketing & Growth Partnerships *Boston, MA*

- Built an end-to-end **ELT data product** for **The Trade Desk**, ingesting TTD APIs into **S3** with schema enforcement, loading to **Redshift** via **Airbyte + Terraform**, and modeling **dimensional dbt marts** (star schema) powering **5 Mode** self-serve reports
- Replaced manual CSV onboarding with an **event-driven SQS, EventBridge, and Kinesis** architecture feeding Everflow data into **Redshift**; implemented **WAP (Write–Audit–Publish)** logic in **Airflow and dbt** with idempotent Everflow reconciliations, cutting onboarding by **90%** and scaling to **150+ partners and $10M+ revenue**
- Delivered **20 stakeholder-facing analytics products** for **30 users**, standardized KPIs, resolved **100+ tickets**, merged **150+ PRs**, and introduced **AI-assisted workflows** (Claude Code, Cursor, **dbt MCP**) to Analytics Engineers
- Refactored two **300GB+ dbt bidding fact models** using **incremental, idempotent pipelines**, optimized **DISTKEYs**, fixed **non-SARGable** joins, and improved partition strategy, reducing **warehouse storage by 100GB** and **runtime by 50%**
- Built **Google SEM reverse ETL pipelines** with **CDC** via **Hightouch**, ingesting **Google Ads APIs via GAQL** into **Redshift** with Airbyte and Airflow; designed **dbt attribution models**, lifting margin by **20%**
- Investigated **vendor-reported missing clicks** using **end-to-end data lineage**, audit logs, and **observability** across **Airbyte, S3, dbt, and Redshift**; reconciled **13 partner networks** in Mode, recovering **$500K** and boosting margin by **25%**
- Deployed **5 production anomaly detection alerts** using **Airflow, dbt tests, source freshness checks**, and **diagnostic dashboards** with Slack alerting, improving **data reliability SLAs** and reducing incidents by **30%**

Lazard **May 2024 – August 2024**

Data Engineer Intern - Financial Advisory Business Technology *New York, NY*

- Migrated a **Monte Carlo simulation** from Jupyter to a **fault-tolerant PySpark pipeline** on **AWS EMR**, orchestrated via **Airflow** with **retryable, idempotent tasks**, reducing runtime by **40%** and productionizing daily revenue forecasts
- Built **5 Streamlit data products** backed by **Alteryx APIs** and **Snowflake**, improving self-service analytics for **30 bankers**
- Designed a **Snowflake analytics model** in **dbt** with standardized transformations, tests, and docs for BI consistency

TechSmith **January 2024 – May 2024**

Business Intelligence Engineer Intern - Marketing & Sales *East Lansing, MI*

- Engineered a **real-time, event-driven pricing pipeline** using **Kafka** to ingest Salesforce data into **Databricks**, where **PySpark MLlib and TensorFlow** processed **5M+ records daily**, improving accuracy by **18%**
- Deployed **dbt models in Snowflake** to unify fragmented datasets, enabling a **single source of truth** for **20+ dashboards**
- Built scalable **ETL pipelines** with **PySpark** and optimized **SQL**, applying **partitioning** and **incremental processing** to reduce turnaround time by **50%**

Corning **January 2023 – May 2023**

Data Analyst Intern - Global Supply Chain *Dallas, TX*

- Applied **analytics and Lean Six Sigma** to deliver operational insights, reducing supply chain waste by **13%**
- Modernized legacy workflows with **Airflow-orchestrated dbt pipelines**, automating **80%** of supplier reconciliation
- Built a scalable **Databricks demand forecasting pipeline** processing **50M+ inventory records** to optimize logistics routing

Projects

CFPB Consumer Complaints Warehouse Pipeline | *Python, dlt, dbt, DuckDB, Prefect, Visivo, GitHub Actions* | GitHub

- Built an **ELT pipeline** using **dlt**, **DuckDB**, and **Prefect** with **Parquet**-staged incremental loads and state management
- Modeled data into a **Star Schema** with **dbt** and delivered **4 Visivo dashboards** powered by **CI/CD** via **GitHub Actions**

NYC Taxi Lakehouse Pipeline | *Dagster, MinIO, Apache Iceberg, Trino, dbt, Docker* | GitHub

- Built a **lakehouse** with **Dagster**, **MinIO**, **Iceberg**, and **Trino**; loaded NYC taxi data via atomic snapshots
- Modeled **silver/gold medallion layers** in **dbt-trino**, delivering **3 gold analytics models** with full asset lineage in **Dagster**

Coinbase Crypto Streaming Pipeline | *Kafka, Spark Streaming, Flink, ClickHouse, FastAPI, Streamlit, Docker* | GitHub

- Streamed real-time **Coinbase WebSocket** crypto and synthetic stock prices through **Kafka** into **Spark Streaming** and **Flink**
- Stored checkpointed windowed aggregations in **ClickHouse** and served live dashboards via **FastAPI** and **Streamlit**

Technical Skills

Business Intelligence & BI as Code: Power BI, Tableau, Looker Studio, Hex, Mode, Visivo, Streamlit, Evidence.dev, Lightdash
Data Warehouse/Lakehouse: Snowflake, Redshift, BigQuery, Databricks, DuckDB/MotherDuck, Iceberg, Delta Lake, Trino
Data Processing & Transformation: PySpark, Apache Spark, Spark Structured Streaming, Alteryx, Kafka, dbt Core
Data Integration: Airbyte, Fivetran, dltHub, Hightouch
Orchestration: Apache Airflow, Astronomer, Prefect, Mage, Dagster, Orchestra
DevOps & Infrastructure: Terraform, Docker, Kubernetes, Git, GitHub, GitLab, Jenkins
Cloud Platforms: AWS (S3, EMR, Glue, Kinesis, SQS, EventBridge), GCP (Dataflow, Cloud Storage)

Education

Texas Christian University — GPA: 3.74 *(Magna Cum Laude)* **Aug 2021 – May 2025**

Computer Systems Analysis & Minor in Mathematics, Fintech *Fort Worth, TX*