Ha Dang Long

Hanoi | longha6104@gmail.com | (+84) 359 716 638 | lh6104 | lh6104 | Portfolio

Summary

Information Systems graduate (Aug 2026) with a strong interest in fintech and in building trustworthy data infrastructure for financial services. Experienced in ETL/batch and CDC pipelines into a Bronze/Silver/Gold data lakehouse (Debezium, Kafka, Spark, Airflow, dbt) and in event-driven ingestion (Python, RabbitMQ, PostgreSQL, MinIO). Focused on verifiable pipelines: reconciliation, idempotency, data-quality checks, and performance tuning (8.64s to 0.66s, 13×).

Experience

Austdoor Group – Data Engineer Intern Aug 2026 – Present

Internal installer-prospecting platform for a referral program (rollout in progress).

- Developed event-driven ingestion components with Python asyncio, RabbitMQ, PostgreSQL, and MinIO, consolidating approved data sources into deduplicated installer prospect records.
- Built URL discovery, deduplication, and Serper/DuckDuckGo failover; a pilot persisted **129 records from 150 URLs**, with SHA-256 verification of **129/129 raw artifacts**.
- Implemented a transactional outbox, retries, and dead-letter queues; observed no lost or duplicate receipts across **18 local SIGKILL test cases**. Fixed a race condition to enforce the **100-record/run limit**.
- Optimized MinIO orphan reporting with batched PostgreSQL queries, reducing lookups from **9,500 to 10** and execution time from **8.64s to 0.66s (13× faster)**.

Beet Innovators – AI Intern Jun 2025 – Aug 2025

- Debugged C++/Qt and OpenCV modules for real-time face detection, recognition, and liveness verification; reproduced defects and performed functional testing under senior guidance.

Projects

Industrial Manufacturing Platform GitHub Oct 2026 – Nov 2026

- Built an end-to-end Bronze/Silver/Gold data lakehouse (S3/Parquet) integrating PostgreSQL snapshots, Debezium CDC, and MQTT–Kafka–Spark Streaming; orchestrated Airflow batch refreshes and dbt marts (**55 models, 75 tests**).
- Reconciled **824,471 source-to-Bronze rows** via count/key-value hashes to verify post-load completeness and accuracy; processed **9.68M measurements** to Gold at ∼**3,551/s** in a batch benchmark; resolved an object-storage OOM by raising its memory cap from **512 MiB to 2 GiB**.
- Implemented durable handoff and idempotent replay (edge spool, bridge inbox, offset ledger), validating **115,200 Kafka envelopes** by SHA-256 and offset parity (final lag 0); handled late events with **15-min watermarks** and a correction queue.
- Ran **14 fault-injection cases** for pipeline monitoring and recovery; a CDC alert fired in **19.8s**, and same-host restore lost **0 rows** in the tested scenario.

Cognitive Traffic Analytics Platform GitHub Mar 2026 – Jun 2026

- In a 2-member team, developed traffic, weather, and news ingestion from REST/RSS/HTML sources and PySpark/Apache Iceberg Bronze–Silver–Gold transformations.
- Implemented Kafka validation, retries, idempotent processing, and dead-letter handling; added data-quality checks and reproducible Docker workflows.

University Alumni Knowledge Graph & GraphRAG Chatbot GitHub Nov 2025 – Dec 2025

- Built a Vietnamese Wikipedia-to-Neo4j pipeline in a 2-member team (**2,100+ entities,** ∼**70,000 relationships**); used Cypher validation and multi-hop retrieval, reaching **75% answer accuracy** on **2,000+ evaluation questions**.

Education

University of Engineering and Technology – Vietnam National University Oct 2022 – Aug 2026

B.S. in Information Systems

Technical Skills

Programming: SQL, Python (asyncio, Pydantic, FastAPI), Java, Linux command line
Data Engineering: ETL/ELT (batch), CDC (Debezium), Spark/PySpark, Kafka, Airflow, dbt, Data Lake/Lakehouse (Apache Iceberg, S3/Parquet), Data Quality
Databases & Messaging: PostgreSQL, MySQL, Neo4j, MinIO/S3, RabbitMQ
Reliability & Tools: Transactional Outbox, Idempotency, Docker, Docker Compose, Git, GitHub Actions, Pytest, Prometheus, Grafana
English: B2 (VNU Tests)