Armaan Thapar

athapar.2017@gmail.com | (914)-486-0735 | armaanthapar.com | GitHub | LinkedIn

Summary: Data Engineer, 4+ years building production batch and analytical systems on AWS, recently designing end-to-end streaming lakehouses (Kafka, Spark Structured Streaming, Delta Lake) for market data. Strength in idempotent ingestion, orchestration, and data-quality reconciliation at scale.

Skills

Languages: Python, SQL
Streaming: Kafka, Spark Structured Streaming, Delta Lake
Warehousing: Redshift, Databricks, Snowflake, BigQuery
Orchestration: Airflow, Prefect, dbt
Cloud & Infra: AWS (Lambda, SQS), Docker, Git, Azure DevOps
Concepts: ETL/ELT, SCD2, Idempotency, Medallion Architecture
AI/ML: Strands, LangChain, LangGraph

Experience

Data Engineer: **Georgia-Pacific** – Atlanta, GA | 05/2022 - Present

- Architected and productionized an analytical platform on AWS Lambda processing operational time-series data for 4 industrial facilities, incorporating 50+ input time series per facility, ad hoc numerical workflows with configurable pipelines using drift-adjusted forecasting and optimization-based parameter inference
- Engineered a staged orchestration system with dependency-aware sequencing, input validation, and telemetry-driven throttling, compressing model deployment time of 10k+ models from ~8–10 months to ~2 months at scale
- Built Python graph-parsing and DFS-based path enumeration tooling to automate model generation and conditional flow analysis across the asset graph
- Designed SQL retrieval and aggregation logic for Strands-based summarization workflows, improving context quality and reducing LLM inference costs on downstream pipelines

Analyst: **Reckitt** – Belle Mead, NJ | 01/2021 – 12/2021

- Built regression models and spectroscopic analysis workflows for raw materials quality data for 1000+ samples/month

Projects

Streaming Market Data Platform | 05/2026 - Present | GitHub | Dashboard

- Built a market data lakehouse ingesting ~37M trades, quotes, and aggregates for 104 equities over a 6.5-hour trading session via Polygon WebSockets, Kafka, Spark Structured Streaming, Delta Lake on Databricks, and Snowflake
- Engineered end-to-end idempotent delivery by coordinating Kafka offset commits with atomic Delta checkpoints and event-specific Silver MERGE keys (window_start for aggregates, trade_id for trades, symbol/timestamp/sequence for quotes), validated by dbt reconciliation marts across grain, value, and row count
- Optimized Snowflake synchronization by replacing row-wise inserts with staged PUT + COPY INTO bulk loading, achieving ~50× faster ingestion on a 7.4M-row trade dataset (~12 min → ~15s)
- Built dbt reconciliation and observability marts comparing streaming OHLCV against batch BigQuery ground truth via composite FIGI, with batch-ingest-lag and session-coverage metrics derived from Gold event timestamps

Equity Data Warehouse | 01/2026 - 05/2026 | GitHub

- Architected batch equity platform ingesting 500K+ daily bars, quarterly financials, dividends, and corporate actions for 104 equities across 20 yrs from Polygon REST APIs into BigQuery via idempotent Parquet-first ingestion via Prefect orchestration
- Designed SCD2 security master with retrospective history by seeding Polygon ticker events (FB→META, MWD→MS) alongside a forward-looking dbt snapshot, enabling point-in-time identity resolution across the full 20-year backfill
- Built analytical dbt marts for TTM financial aggregation, split-adjusted pricing, valuation metrics, and factor-based ranking workflows across corporate actions and evolving entity identities
- Engineered resilience with HTTP retry/backoff, per-symbol isolation, and batched BigQuery loads (5 jobs vs 520 per-symbol writes); 43 dbt tests enforce grain, SCD2 integrity, and TTM completeness

Education

The Cooper Union, New York, NY – Chemical Engineering, **MEng**, (Dec 2020); **BEng.** (May 2018)