

Sarthak Mistry

✉ sarthakmistry44@gmail.com ☎ +1 (930) 904-4600 [in linkedin](#) [github](#) [portfolio](#) 📍 Open to relocation

Summary

Data Engineer with 2+ years of experience building batch and real-time pipelines (Snowflake, Spark, Kafka, Airflow) that drives product and business decisions in financial services. Skilled in dimensional data modeling (Data Vault 2.0, star schema), data quality frameworks, and cross-functional collaboration with analysts, product, and compliance teams. Seeking product-analytics-focused data engineering roles where pipeline design directly shapes user experience at scale.

Professional Experience

Data Engineer 1 — *Piramal Capital & Housing Finance, Mumbai, India* Jul 2022 – Jul 2024

- Built batch ETL pipelines ingesting 5M+ daily records from Salesforce, MongoDB, SQL Server, and REST APIs into Snowflake using Spark and Airflow, powering dashboards used by 30+ analysts and business stakeholders for loan portfolio monitoring.
- Designed and automated a 150-feature regulatory data mart in Snowflake (AWS Lambda, S3), improving compliance team's regulatory reporting turnaround by ~40% and directly supporting quarterly RBI audit submissions.
- Reduced PII exposure risk by building an automated detection and masking pipeline using regex patterns and Snowflake dynamic masking, achieving 93% precision across 5M+ customer records in production audits.
- Deployed a real-time CDC pipeline (Debezium, Kafka, Spark Streaming) feeding Grafana dashboards for live loan disbursement tracking, enabling ops managers to identify and resolve funding bottlenecks within minutes.
- Collaborated with chief economist to instrument economic indicator scraping (Python, Selenium), reducing manual data collection effort by 73% and feeding predictive models for loan default risk.

Data Engineering intern — *NSEIT, Mumbai, India* Oct 2021 – Mar 2022

- Engineered a CI/CD pipeline for 20+ dbt models in PostgreSQL, automating pre-aggregated data modeling for analytics team and reducing Kibana dashboard query latency by 40%.
- Implemented automated dbt schema validations and data quality tests for candidate demographic datasets, reducing data ingestion errors by 25% and improving analyst confidence in downstream reporting.

Projects

E-commerce Batch Data Pipeline — *Airflow, dbt, Snowflake, Great Expectations* 🔗

- Built an end-to-end analytics pipeline on Snowflake with a 4-layer medallion architecture (RAW → Staging → Data Vault 2.0 → Star Schema) across 17 dbt models, processing TPC-DS SF100 benchmark data (~100GB, 25 tables).
- Implemented 50+ data quality expectations across 7 Great Expectations suites with a quarantine-not-hard-fail pattern, routing bad records to a Snowflake triage table for analyst review.
- Designed a Data Vault 2.0 layer (3 hubs, 2 satellites, 1 link) with MD5 hash-diff change detection and SCD Type 2 via dbt snapshots; built a downstream OBT with RFM segmentation, MoM growth, and rolling revenue windows.
- Orchestrated 2 Airflow DAGs via Astro CLI and Astronomer Cosmos, rendering each dbt model as an individual task for per-model retry and observability.

Blockchain Streaming Data Pipeline — *Spark, Trino, Hudi, Kafka* 🔗

- Designed a self-managed 3-node Kappa architecture cluster (Hadoop, Kafka, Spark, Trino) processing ~500 live Bitcoin transactions/min from Blockchain.com's WebSocket feed.
- Trained a KMeans clustering model (SparkML) on BigQuery's public Bitcoin dataset to score transaction anomalies; deployed as a Spark Structured Streaming job writing results to Apache Hudi tables.
- Integrated Trino with Hive Metastore as the query layer, enabling low-latency SQL analysis of streaming results for ad-hoc investigation.

Technical Skills

Languages & Query: Python, SQL, Spark (PySpark), JavaScript, Rust, Bash

Data Engineering: Snowflake, dbt, Apache Airflow, Apache Kafka, Apache Spark, Debezium CDC, Trino/Presto, Apache Hudi

Data Modeling & Quality: Star schema, Data Vault 2.0, SCD Type 2, dimensional modeling, Great Expectations, data lineage

Infrastructure & Tools: AWS (S3, Lambda, EMR, Glue), Docker, Kubernetes, Terraform, Talend, Git, Astro CLI

Visualization & BI: Grafana, Power BI, Streamlit, ELK/Kibana; dashboard design for operational and analytical use cases

Education

Indiana University, Luddy School of Informatics, Computing and Engineering

May 2026

Master of Science in Data Science

University of Mumbai

May 2022

Bachelor of Engineering in Electronics and Telecommunication Engineering