

PROFESSIONAL SUMMARY

Data Engineer with hands-on experience building production ETL/ELT and streaming pipelines for healthcare and fintech clients across Databricks, Azure, and AWS. Proficient in PySpark, dbt, Snowflake, and Delta Lake for batch and real-time workloads. Strong background in data governance – including dynamic data masking, tag-based lineage, and schema drift detection – alongside Airflow orchestration and CI/CD automation. Databricks Certified Data Engineer Associate.

TECHNICAL SKILLS

Languages: Python, SQL, Java, Scala

Orchestration & Transformation: Apache Airflow (MWAA), dbt, IBM DataStage, Azure Data Factory, AWS Glue

Cloud – AWS: S3, Lambda, Redshift, Athena Spectrum, EMR, Step Functions, EC2

Cloud – Azure: Data Factory, Stream Analytics, Databricks, Azure DevOps (CI/CD)

Processing & Streaming: PySpark, Spark SQL, Apache Kafka, Apache Flink

Data Platforms: Databricks, Snowflake, Delta Lake, Microsoft Fabric

Data Stores: PostgreSQL, MongoDB, DuckDB

Data Modeling: Star schema, Snowflake schema, dimensional modeling

DevOps & Tools: Docker, Git, Azure DevOps, Jira, Linux, Airbyte, Terraform, Scrapy

BI & Visualization: Power BI, Tableau

EXPERIENCE

Data Engineer

Dubizzle Labs

March 2026 – Present

Lahore, PK

- Optimized SQL query performance across core data pipelines serving the Classifieds, Motors, and Properties verticals, cutting execution time by 30% and reducing downstream compute cost.
- Led migration of a legacy source system to a new platform, redesigning ingestion logic to ensure data integrity and zero-downtime cutover across high-volume listing data.
- Built and optimized AWS Athena workflows for ad-hoc analytics over large-scale, partitioned datasets spanning multiple business verticals.
- Improved CDC pipeline reliability by enhancing WAL (Write-Ahead Log) binlog parsing logic with schema evolution support, preventing pipeline breakage from upstream schema drift.

Associate Data Engineer

Alpha Bridge – Healthcare and Fintech Clients

February 2025 – February 2026

Lahore, PK

- Built Spark Declarative Pipelines in Databricks with Delta Lake and expectations-based data quality checks, reducing pipeline failure rates and enabling medallion-architecture data flow.
- Designed and implemented Snowflake dynamic data masking policies controlled via object tags; built a stored procedure to trace tag and data lineage across schemas, improving the team's masking workflow efficiency by 20%.
- Automated column-level tagging in dbt through custom macros, eliminating manual tagging for newly added columns and ensuring consistent governance across all downstream models.
- Migrated legacy ActiveBatch scheduled workflows to Apache Airflow (MWAA), improving observability, retry logic, and SLA adherence across production DAGs.
- Implemented schema drift detection in IBM DataStage to flag structural changes and prevent silent pipeline failures across fintech ETL jobs.
- Reduced streaming data latency from 10 to 7 minutes by tuning Azure Stream Analytics windowing functions and optimizing ADF pipeline triggers.
- Implemented CI/CD for Azure Data Factory using Azure DevOps, automating ARM template deployments across dev, staging, and production environments.

Software Engineer (Backend)

Sunsprinkle

March 2023 – December 2023

Lahore, PK

- Built Django REST backend with token-based authentication and role-based access control for a course management platform, and developed Scrapy crawlers with proxy rotation for structured data collection at scale.

PROJECTS

CMS Synthetic Patient Data Pipeline

Pipelines

Synthea, Apache Airflow, Databricks Asset Bundles (DABs), Lakeflow Declarative

- Built a pipeline to generate high-volume, CMS-compliant synthetic patient data using Synthea, orchestrated with Airflow and deployed via Databricks Asset Bundles (DABs) with Lakeflow Declarative Pipelines for ingestion and transformation.
- Tuned Synthea's module probability weights and disease-progression distributions to produce more realistic, real-world-representative patient cohorts instead of relying on default synthetic distributions.
- Identified and handled unbillable medical codes in generated claims data, adding validation logic to improve billing-code accuracy and downstream data usability.

Schema Drift Detection and Evolution Engine

Airflow, DuckDB, S3, Email Notifications

- Built an automated schema drift detection pipeline using Airflow and DuckDB that monitors source schemas, sends email alerts on structural changes, and conditionally evolves or archives data to S3 based on human-in-the-loop email approval.

NYC Taxi ELT Pipeline

Airflow, AWS Lambda, S3, Snowflake, dbt

- Designed a production-grade ELT pipeline: Lambda extracts and lands raw data in S3, Snowflake handles warehousing, and dbt manages transformations with built-in schema drift detection and automated testing.

E-Commerce ETL Pipeline – [GitHub](#)

Python, Snowflake, Docker, AWS EC2

- Dockerized ETL pipeline deployed on EC2 that extracts, transforms, and loads e-commerce data into Snowflake with incremental loading, data validation, and automated scheduling.

CERTIFICATIONS

Databricks Certified Data Engineer Associate

February 2026

Databricks – [Verify Credential](#)

EDUCATION

Bachelor of Science in Data Science

July 2025

University of Central Punjab – CGPA: 3.34

Lahore, PK