

Shreyas Mysore Sudesh

Data Engineer



518-956-5361



shreyasmsudesh28@gmail.com



LinkedIn



GitHub



Portfolio

SUMMARY

Data Engineer with 4 years of experience engineering high-reliability ETL/ELT pipelines, reducing data incidents by 60%, achieving 97% uptime across 10M+ daily records, and optimizing Spark workflows from hours to minutes at enterprise scale.

SKILLS

Methodologies: Agile, SDLC, Waterfall

Programming & Scripting: Python, SQL, Bash, PySpark

Data Warehousing & Databases: Snowflake, Amazon Redshift, PostgreSQL, Azure SQL DB, BigQuery

Pipeline Orchestration & Streaming: Apache Airflow, dbt, Dagster, Apache Kafka, Spark Streaming

Big Data Processing: Apache Spark, Databricks, EMR, Unity Catalog

Data Modeling & Architecture: Star Schema, SCD Type 2, Medallion Architecture

Data Quality & Observability: Great Expectations, data validation, pipeline monitoring, SLA management

Cloud Platforms: AWS (S3, Glue, Lambda, EMR, Redshift), Azure (Azure Synapse, Azure Data Factory, ADLS)

PROFESSIONAL EXPERIENCE

Data Engineer | BNY, New York

January 2026 - Current

- Reduced data incidents 60% engineering automated quality monitoring using **Python**, **Great Expectations**, and **Airflow** with **dbt** tests, enforcing CI/CD checks via **GitHub Actions** across analytics and ML training workflows.
- Delivered 99+% pipeline uptime engineering production **ETL/ELT** workflows processing 10M+ daily records using **Python**, **SQL**, **Airflow**, and **GCP Dataflow**, powering executive BI reporting and downstream ML feature pipelines.
- Engineered 50+ ML-ready features using **Databricks** and **PySpark** for fraud risk models, orchestrating pipelines with **Dagster** and training models directly in **BigQuery ML**, ensuring reproducible lineage for retraining.
- Built real-time fraud detection infrastructure using **Kafka** and **PySpark**, delivering low-latency event pipelines that scored 10M+ daily transactions for deployed ML anomaly detection models in production.
- Engineered optimized **BigQuery** semantic layers and aggregated data models consumed by **Power BI**, supporting fraud risk reporting and executive KPI visibility, reducing ad-hoc data requests to engineering by 40%.

Data Engineer | Tredence Inc, India

March 2023 – August 2024

- Engineered 100+ **Azure Data Factory** pipelines ingesting 100+ GB/day of retail, customer, transaction, and tender data from **Oracle** into **Databricks Delta Lake**, sustaining 150+ curated tables at 98% success rate.
- Developed Bronze, Silver, and Gold **Delta Lake** layers using **PySpark** and **Spark SQL**, landing curated data in **Azure Data Lake Storage (ADLS)** Silver zone, standardizing transformations for enterprise-wide analytics.
- Implemented **CDC-aware** incremental ingestion and **Delta MERGE upsert** patterns in **PySpark**, building **SCD Type 2** dimensional models to preserve historical customer and business attributes, improving reporting accuracy for trend analysis.
- Built **SQL** and **PySpark** validation and reconciliation checks across curated tables, partnering with **BI** stakeholders to define data contracts, achieving 99% on-time SLA and 30% faster resolution, improving reporting trust.
- Delivered 20+ **Power BI** and **Tableau** dashboards on Gold-layer datasets, enabling cross-functional teams to monitor customer activity, channel performance, tender trends, and operational KPIs across business functions for enterprise-wide leadership.
- Containerized Spark ETL jobs using **Docker** and deployed via **Kubernetes (AKS)**, enabling automated **CI/CD** pipelines that improved deployment consistency and reduced manual release effort across environments.

Data Engineer | Citus Infotech, India

April 2021 – February 2023

- Led the design and implementation of a centralized data lake using **AWS S3** and **PySpark**, integrating multiple data sources while leveraging **Python** for streamlined quality checks, resulting in 25% reduction in data preparation time.
- Developed a robust monitoring system using **Hadoop** and **Power BI**, enabling real-time visualization of pipeline metrics and alerts, reducing system downtime by 20% through issue detection.
- Collaborated with Data Scientists to optimize **data models** and enhance data accessibility while implementing robust security measures and **AWS** infrastructure optimizations, resulting in improved system performance and reduced operational costs by 20%.
- Architected and deployed a scalable data warehouse solution using **Snowflake** and **Python**, efficiently processing 10GB daily data while maintaining optimal query performance, enabling self-service analytics for 7 business users.

EDUCATION

Master of Science in Business Analytics | University at Albany, SUNY, US

May 2026

ACADEMIC PROJECTS

NYC 311 Service Demand Data Pipeline: Built a Terraform-provisioned BigQuery analytics pipeline processing 4.4M+ NYC 311 service requests with standardized ingestion, transformation, modeling-ready tables, and an XGBoost predictive model achieving $R^2 = 0.786$.

Snowflake Analytics Engineering Pipeline: Created a dbt and Snowflake analytics engineering project over 5M+ subscription and attribution records, producing 25 analytics models and 40+ dbt tests for schema integrity, null checks, uniqueness, accepted values, and referential integrity.

Real-Time Streaming Pipeline: Built a Kafka-to-Delta Lake streaming pipeline processing approximately 10,000 events/min across 3 topics with checkpointing, idempotent writes, Bronze/Silver design, and sub-5-second observed latency.

CERTIFICATIONS

Databricks Certified Data Engineer Associate

Great Learning– Data Analytics Program