Dhruvraj Singh Rathore

Bryan, Texas | dhruvrajrathore2011@gmail.com | dhruvrajsinghrathore.github.io/Portfolio/ | linkedin.com/in/dhruvrajsingh/ | 7372061179

SUMMARY

Data Engineer with **2+ years specializing in ETL and ELT architecture** on AWS and Snowflake, with **expertise in Spark** performance tuning and **SQL-based analytical modeling**. Delivered reliable, low-latency data pipelines with production data quality enforcement using Airflow and dbt.

EDUCATION

MS in Data Science, Texas A&M University, College Station, Texas, USA | GPA: 4.0 — Aug 2024 - Dec 2025
BS in Computer Science, SRM Institute of Science and Technology, India | GPA: 3.8 — Jul 2018 - May 2022

TECHNICAL SKILLS

Data Engineering: Python, SQL, PySpark, Apache Spark, Apache Airflow, Kafka, dbt, ETL/ELT, Data Modeling
Cloud & Databases: AWS (S3, Glue, Redshift, Lambda, Athena), Snowflake, Delta Lake, PostgreSQL, MongoDB, Redis
Tools & Orchestration: Docker, Kubernetes, Git/GitHub, CI/CD (GitHub Actions), Tableau, AWS Quicksight

WORK EXPERIENCE

Data Engineer, Draup Business Solutions, Bangalore, India — Dec 2022 - Jun 2024

- **Reduced Spark ETL runtime by 2 hours** by optimizing partitioning and schema configuration across 50+ TB S3 parquet data.
- **Improved query execution time by 40%** by redesigning legacy **Snowflake** and **Amazon Redshift** analytical warehouse models using optimized clustering, distribution keys, and schema structure for reporting workloads.
- Improved MongoDB read throughput by **A/B testing** indexing and sharding strategies for production monitoring workloads.
- **Collaborated with analytics teams** to design and implement a serverless data pipeline using **AWS** Lambda and DynamoDB to support on-demand analytical queries on S3 data, **reducing client data delivery time by 50%.**
- **Led cross-functional data quality initiative** by developing **PySpark** validation checks and orchestrating daily schema, null, and freshness tests in Airflow across 200M+ records, **reducing QA escalations by 70%.**

Data Analyst, HighRadius Corporation, Hyderabad, India — Jul 2022 - Nov 2022

- **Accelerated monthly financial report generation time by 40%** by introducing materialized views and optimizing aggregation SQL queries in Snowflake.
- Built data validation checks in SQL to verify payment records before reporting, catching data quality issues early and **reducing downstream corrections by 60%**.
- Designed Tableau dashboards highlighting customer payment risk and aging trends, **reducing manual collection effort by 60% and improving cash recovery by 20% across 20+ collection teams.**

PROJECTS

Real-Time Website Behavior Analytics | GitHub-Link — Jan 2026 - Feb 2026

- Created **Kafka-based streaming pipeline** to process real-time clickstream events with 1-minute window aggregations, storing results in MinIO object storage for analytics and MongoDB for live monitoring.
- Developed **threshold-based anomaly detection** flagging users exceeding 50 events per minute, enabling sub-minute detection of suspicious bot or fraud activity.
- **Deployed the full system using Docker** across 8 integrated services, reducing local environment setup time by **90%** and ensuring consistent service orchestration with automated health checks and topic initialization.

Scalable Analytics Lakehouse with Spark | GitHub-Link — Dec 2025 - Jan 2026

- **Engineered scalable analytics lakehouse platform in Spark** and Delta Lake using **medallion architecture** to support ingestion, transformation, and analytics-ready datasets with rule-based validation.
- **Reduced analytics query latency by 85%** by materializing gold layer aggregates to MongoDB for low-latency API access.
- Orchestrated production batch workflows by deploying containerized Spark jobs via Docker Compose, Airflow, and Kubernetes CronJobs with retry logic, **achieving 99% job success.**

Order Analytics Pipeline | GitHub-Link — Sept 2025 - Oct 2025

- Built end-to-end analytics pipeline using **Airflow** and **dbt** to orchestrate data ingestion, transformation, and loading into Snowflake with layered models, **enabling automated daily reporting refresh**.
- Implemented dbt tests for schema validation and enforced dimensional modeling best practices across fact and dimension tables, catching **95%+ data quality issues pre-production.**
- Containerized the pipeline with Docker and deployed on Astronomer Airflow runtime with centralized monitoring, **reducing deployment time from 2 hours to 15 minutes and enabling reproducible releases.**