RONIT ROY

857-396-6089 | Boston, MA, 02119 | ronitr300701@gmail.com | LinkedIn | github.com/ronitroy30 | ronitroy.live

Summary

Data Engineer with hands-on experience building batch and real-time pipelines on AWS and Spark. Skilled in SQL optimization, data modeling, and orchestration (Airflow, Kafka, dbt). Experienced in implementing data quality checks, monitoring/alerting, and cost optimization to ensure reliable, production-ready pipelines.

Experience

Saayam For All Remote

Data Engineer *Mar. 2025 – Present*

- Built AWS-based ETL pipeline ingesting IRS EO BMF and ProPublica API datasets , reducing manual curation by 80% and enabling public search across six aid categories.
- Configured CloudWatch alarms and Lambda auto-retries, achieving less than 5 min recovery for failed jobs.
- Collaborated with analytics team to integrate datasets into Redshift dashboards for category-level insights.

GLOB S Research Lab Boston, Massachusetts

Research Assistant *Oct. 2023 – May. 2024*

- Designed Airflow DAGs and PostgreSQL schema changes, reducing research dataset turnaround from 2 days to 1.5 days (25%).
- Increased NLP pipeline accuracy to 95% by containerizing models in Docker and automating CI/CD with Jenkins.
- Implemented AWS Glue Catalog lineage to ensure transparent dataset traceability across projects.

HighRadius Bangalore, India

Machine Learning Intern *Jan. 2022 – Apr. 2022*

- Built CNN-based fraud detection model (87% recall, 95% precision) on 50K transaction dataset.
- Deployed Flask APIs in Docker containers; set up monitoring alerts for API downtime, improving availability to 99.9%.
- Automated ETL workflows between Python and Snowflake, reducing prep time by 40%.

Projects

Spotify Data Pipeline *(Python, Snowflake, AWS, Airflow)*

- Built ETL pipeline processing 100K+ daily records using AWS Lambda + Snowpipe.
- Reduced analyst prep time from 4 hours to 45 minutes via automated transformations.

Retail Data Lake *(Spark, Delta Lake, AWS S3)*

- Designed Delta Lake-based architecture on AWS handling terabyte-scale retail data.
- Used partitioning + Z-ordering to reduce query costs by 30% and optimize analyst workflows.

E-commerce Recommendation System *(Python, Flask, TensorFlow, PostgreSQL)*

- Developed a collaborative filtering-based recommendation system using TensorFlow, improving user engagement by 20%.
- Deployed API endpoints using Flask for seamless integration with front-end systems.
- Optimized PostgreSQL queries for real-time recommendation generation.

Advanced ETL for Retail Analytics *(Airflow, AWS, Kafka, Snowflake, Redshift)*

- Constructed a dynamic ETL pipeline handling 500K+ records/hour using Apache Kafka and Snowflake.
- Enabled near real-time analytics through AWS Redshift warehouse optimization.
- Implemented automated error handling in Airflow to ensure data consistency.

Technical Skills

Programming: Python, SQL, R, Java, Shell, Scala
Data Engineering: ETL/ELT pipelines, Data transformation, Data ingestion, Data validation, Schema design
Cloud & Big Data: AWS (S3, Lambda, Glue, Redshift), Spark, Hadoop, Big Query
Orchestration & Tools: Airflow, Docker, Git, Kafka, dbt, Kubernetes, Jenkins, Git
Databases & Warehousing: PostgreSQL, MySQL, Snowflake
Monitoring & Maintenance: : CloudWatch,Prometheus, Grafana
Visualization & Reporting: Tableau, Power BI, AWS QuickSight

Education

Boston University Boston, MA

Masters of Science in Applied Data Analytics *Sept. 2023 – Dec. 2024*

- Coursework: Advanced Machine Learning, Advanced Database Management, Data Mining

SRM Institute of Science and Technology Chennai, India

Bachelor of Technology in Computer Science *Jul. 2019 – Apr. 2023*

- Coursework: Probability and Statistics, Data Structure and Algorithms, Data Visualization