SOHITH SAI M

New York, NY | sohithjobz@gmail.com | +1 716 226 8430 | LinkedIn | GitHub

PROFESSIONAL SUMMARY

Data Engineer with 3+ years of experience designing and operating **scalable ETL/ELT pipelines**, distributed processing workflows, and analytics-ready datasets using **Python, SQL, Spark, and AWS**. Proven track record ingesting **50+ GB/batch** data, improving data quality and integrity, and optimizing pipeline performance (latency **35–45%**, runtimes **40–55%**) with strong monitoring, reliability, and documentation practices. Collaborative partner to analytics and data science teams to deliver trusted data foundations for reporting and ML workloads.

EDUCATION

University at Buffalo, Buffalo, NY — Masters in Engineering Science (Data Science) **Aug 2024 – Jan 2026**
Focus Areas: Data Engineering, Machine Learning, Databases, Cloud Computing

SRM University–AP, India — Bachelors in Technology (Computer Science Engineering) **May 2024**
Relevant Coursework: Data Structures, DBMS, Big Data, Machine Learning

EXPERIENCE

Data Engineer **Jul 2021 – Jul 2024**
Mrdotsolutions Pvt. Ltd

- Designed and implemented scalable **AWS-native data pipelines** using S3, Glue, Lambda, EMR, and Airflow to ingest and transform **50+ GB per batch** of structured and semi-structured data for analytics and reporting.
- Built and optimized **ETL/ELT workflows** in Python and SQL, implementing dimensional models, partitioning strategies, and query tuning to reduce analytics latency by **35–45%**.
- Developed **Spark-based distributed processing** for multi-million-row datasets, optimizing partitioning, caching, and execution plans to reduce runtimes by **40–55%** and improve cluster utilization.
- Implemented **data validation, monitoring, logging, alerting, and failure recovery**, improving reliability to **>99.5% SLA** while optimizing compute/storage costs and enforcing access controls.
- Performed root-cause analysis on pipeline failures and bottlenecks, reducing downtime by **30%** and eliminating recurring incident patterns.
- Partnered with analytics and data science teams to deliver **ML-ready and analytics-ready datasets**, accelerating experimentation cycles and improving training/evaluation efficiency.
- Documented pipeline architecture, data models, and operational runbooks to improve maintainability and handoffs.

PROJECTS

NYC Taxi Data Analysis & Dashboard

- Built a scalable ingestion and processing pipeline handling **50+ GB** of raw trip data, cutting preprocessing time by **40%** via optimized I/O and vectorized Python transformations.
- Developed production-grade ETL workflows to clean, transform, and aggregate multi-million-row datasets, enabling low-latency analytics for reporting and modeling.
- Built interactive dashboards with dynamic filtering and geospatial visualization, consistently achieving **sub-300ms** query latency for end users.

Amazon USA Sales Database & Analytics

- Designed a normalized **PostgreSQL** schema supporting millions of records, enforcing referential integrity and efficient analytical querying.
- Implemented batch ingestion and optimized SQL queries (joins, aggregations, indexing), reducing execution time by **50–60%**.
- Built a **Streamlit** analytics application for self-service reporting, improving business user access to insights by **20%**.

TECHNICAL SKILLS

Programming: Python, SQL, Java, JavaScript, C++
Data Engineering: ETL/ELT, Data Pipelines, Data Modeling, Data Quality, Data Validation, Data Governance
Big Data & Systems: Spark, PySpark, Hadoop, Hive, Snowflake
Cloud: AWS (S3, Glue, Lambda, EMR, Athena, Redshift), Azure Data Factory (Exposure)
Databases & BI: PostgreSQL, Power BI, Tableau, Streamlit, Excel
ML & Frameworks: Pandas, NumPy, TensorFlow, PyTorch, MLflow
Tools & Methods: Docker, Kubernetes, Git, CI/CD, SDLC, Agile

CERTIFICATIONS

AWS Certified Data Engineer – Associate **Dec 6, 2025**