NAVEEN VAYILAPALLI

Data Engineer | 3.9 YOE | PySpark • Airflow • Databricks • AWS • Delta Lake

✉ naveenvayilapalli007@gmail.com | 📞 +91 - 8639327917 | 📍 Hyderabad, India | LinkedIn | Portfolio

Professional Summary

Data Engineer with 3.9 years of experience designing, building, and optimizing scalable batch and near-real-time data platforms using PySpark, Spark SQL, and Databricks. Specialized in developing enterprise ETL/ELT pipelines, implementing Delta Lake Medallion Architecture (Bronze–Silver–Gold), and orchestrating workflows via Apache Airflow. Proven expertise in processing 10M+ records daily, performance tuning Spark workloads, and delivering analytics-ready datasets for BI and reporting. Experienced in AWS cloud deployments and Terraform-driven infrastructure automation for production-grade data platforms.

Technical Skills

Big Data: Apache Spark, PySpark, Spark SQL, DataFrame API | Databases: PostgreSQL, MySQL, MongoDB, Delta Tables
Data Engineering: ETL/ELT Pipelines, Incremental Loads, Change Data Capture (CDC), Data Quality, Data Transformation
Streaming: Structured Streaming, Watermarking, Window Aggregations (Basics)
Data Lakehouse: Delta Lake, Medallion Architecture, ACID Transactions, Time Travel, Schema Evolution
Orchestration: Apache Airflow, DAGs, Sensors, XCom/XComs, TaskFlow API, SLAs
Data Warehousing: Star Schema, Snowflake Schema, SCD Type 1 & 2, Dimensional Modelling, Data Modelling
Cloud & Platform: AWS S3, EC2, IAM, VPC, Databricks Unity Catalog and Workflows
Infrastructure: Terraform (Infrastructure as Code), Environment Provisioning
Programming: Python, Pandas, REST API Integration, OOPs
Optimization: Partitioning, Broadcast Joins, AQE, Skew Handling, Caching, Shuffle Optimization
BI & Analytics: Power BI, DAX, KPIs, Row-Level Security (RLS)

Professional Experience

Tata Consultancy Services — Data Engineer / PySpark Developer **Mar 2022 – Aug 2025**

- Designed and developed scalable PySpark-based ETL pipelines processing 10M+ records daily from SAP MM, flat files, and relational databases into Delta Lake. Implemented Bronze–Silver–Gold Medallion Architecture transforming raw data into curated, analytics-ready datasets with strong data quality controls.
- Executed ETL workloads on Databricks leveraging Delta Lake optimizations and processed datasets up to 500GB+ per pipeline execution. Architected scalable data Lakehouse platforms on Databricks supporting enterprise analytics workloads.
- Built incremental ingestion frameworks and optimized distributed Spark workloads using partitioning, broadcast joins, caching, and Adaptive Query Execution, reducing pipeline runtime by 40%. Engineered SCD Type-2 dimensional models enabling historical tracking for supply chain and demand forecasting analytics.
- Developed and orchestrated 20+ production-grade Apache Airflow DAGs leveraging sensors, XCom/XComs, retries, SLAs, and automated failure alerting, reducing operational delays by 60%. Implemented monitoring, logging, and alerting frameworks for production data pipelines ensuring SLA adherence.
- Wrote and optimized complex SQL queries using CTEs, window functions, indexing, and query tuning techniques improving query performance by 55%. Designed and implemented enterprise Star and Snowflake schemas supporting large-scale analytical and reporting workloads.
- Automated cloud infrastructure provisioning using Terraform (EC2, S3, IAM, VPC), reducing environment setup time from 3 hours to 15 minutes. Developed interactive Power BI dashboards with custom KPIs, drill-through capabilities, and DAX-based time intelligence supporting 150+ business stakeholders.

Environment: PySpark, Spark SQL, Databricks, Delta Lake, Apache Airflow, AWS, Terraform, SQL, Power BI

Certifications

- Deep Learning Specialization — Coursera
- IBM Data Science Professional Certificate - Coursera
- Machine Learning Specialization – Coursera

Projects

Real-Time Streaming Data Pipeline

Designed real-time ingestion framework using Kafka and Spark.

- Implemented watermarking for late data handling.
- Managed checkpointing and state storage for fault-tolerant processing.
- Built window aggregations for KPI reporting.
- Stored curated data in Delta Lake.

Tech Stack: Kafka, PySpark, Structured Streaming, Delta Lake, Watermarking

PySpark Airflow ETL Pipeline

- Built production-grade ETL pipelines using PySpark and Delta Lake.
- Implemented Bronze–Silver–Gold Medallion Architecture with data quality checks.
- Developed YAML-driven configurable ingestion frameworks.
- Orchestrated Spark jobs via Airflow TaskFlow API with SLA monitoring.
- Optimized pipelines using partitioning and incremental processing.

Tech Stack: PySpark, Airflow, Delta Lake, YAML
GitHub: github.com/NaveenVayilapalli007/pyspark-airflow-etl-pipeline

Uber Trips Analytics Dashboard

Built end-to-end analytics dashboards tracking trip volume, revenue, and driver KPIs.

- Designed dimensional data models for scalable reporting.
- Developed drill-through and time intelligence reports using DAX.
- Created paginated PDF reports via Power BI Report Builder.
- Automated refresh and distribution using Power BI Service.

Tech Stack: Power BI, SQL, DAX

Education

B.Tech — Computer Science — Rajiv Gandhi University of Knowledge Technologies 06/2015 - 08/2021
CGPA: 9.0 / 10 (85.5%)