

# NAVEEN VAYILAPALLI

DATA ENGINEER | PYSPARK | AIRFLOW | SQL | TERRAFORM | PYTHON

## CONTACT

- +91 8639327917
- naveenvayilapalli007@gmail.com
- Hyderabad, India 501510

## SKILLS

- Data Engineering & Analytics**
  - PySpark, Spark SQL, Databricks
  - Delta Lake, Medallion Architecture
  - ETL / ELT Pipelines, SCD Type 1 & Type 2
  - Star & Snowflake Schema
  - Data Modelling & Performance Tuning
- Orchestration**
  - Apache Airflow
  - Dynamic DAG Generation
- Databases**
  - PostgreSQL, MySQL, Delta Lake
  - MongoDB
- Cloud & DevOps**
  - AWS: S3, EC2, IAM, VPC
  - Terraform (Infrastructure as Code)
- Programming & Tools**
  - Python, Pandas, OOPs, REST API Integration
  - Linux, Shell Scripting, Git & GitHub
- FastAPI (Basics)**
- Business Intelligence**
  - Power BI, DAX, KPIs, RLS, Power BI Service



## PROFESSIONAL SUMMARY

Data Engineer with 3+ years of experience designing, building, and optimizing scalable ETL pipelines using PySpark, Airflow, SQL, and Databricks. Proven expertise in implementing Delta Lake Medallion Architecture (Bronze-Silver-Gold), SCD Type-2 models, and performance-optimized data transformations. Hands-on experience with AWS cloud services and Terraform (IaC) for infrastructure automation. Strong at translating business requirements into reliable, analytics-ready data platforms.



## WORK EXPERIENCE

03/2022 - 08/2025

### Tata Consultancy Services Hyderabad, India

Data Engineer, Pyspark developer

- Designed and developed 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 to ensure data quality, scalability, and analytics readiness.
- Optimized Spark jobs using efficient joins, partitioning, caching, and broadcast strategies, reducing pipeline runtime by 40%.
- Built SCD Type-2 dimensional models to support historical tracking for supply chain and forecasting analytics.
- Developed 20+ Apache Airflow DAGs with sensors, XComs, retries, SLAs, and failure alerts; reduced operational delays by 60%.
- Wrote advanced SQL queries using CTEs, window functions, indexing, and query tuning, improving query performance by 55%.
- Designed and optimized Star and Snowflake schemas for analytical workloads.
- Automated cloud infrastructure provisioning using Terraform (EC2, S3, IAM, VPC), reducing setup time from 3 hours to 15 minutes.
- Developed Power BI dashboards with custom KPIs, drill-throughs, and DAX-based time intelligence, supporting 150+ business stakeholders.



## EDUCATION

06/2015 - 08/2021

### Rajiv Gandhi University Of Knowledge Technologies, Nuzvid

Graduated in Computer Science with 9.0 GPA

Member of National Social Service [N.S.S], 2016 to 2020 and participated in related social awareness camps and volunteered in Swach Bharath activity etc.



## CERTIFICATIONS

- Deep learning specialization**  
<https://www.coursera.org/account/accomplishments/specialization/C7C4D3G382X5>
- IBM Data Science specialization**  
<https://www.coursera.org/account/accomplishments/professional-cert/QX54DPC432JM>
- Machine learning specialization**  
<https://www.coursera.org/account/accomplishments/specialization/2MFETA983ZJD>



## PROJECTS

### PySpark Airflow ETL Pipeline | GitHub

- Built a production-grade ETL pipeline using PySpark, Delta Lake, and Airflow TaskFlow API.
- Implemented Bronze–Silver–Gold architecture with data quality checks and analytics-ready aggregations.
- Designed config-driven pipelines with YAML and path-safe resolution for portability.
- Orchestrated Spark jobs using Airflow with runtime-safe resource initialization.
  - Tech: PySpark, Airflow, Delta Lake, YAML

Github Repo Link : <https://github.com/NaveenVayilapalli007/pyspark-airflow-etl-pipeline>

### Uber Trips Analytics Dashboard

- Developed an interactive analytics solution to monitor trip volume, revenue, trip status, and driver performance.
- Created paginated reports for PDF exports using Power BI Report Builder.
- Automated report refresh and distribution through Power BI Service.
  - Tech Stack: Power BI, SQL, DAX, Time Intelligence

### COVID-19 Analytics Dashboard

- Built automated data ingestion pipelines using Python and web scraping to collect live COVID-19 data.
- Performed exploratory data analysis and visualized real-time trends.
- Delivered dashboards tracking case growth, recovery, vaccination, and regional spread.
  - Tech Stack: Python, Pandas, Matplotlib, Bokeh, Requests, BeautifulSoup

### Signature Forgery Detection (Academic Project)

- Developed a deep learning-based system to classify genuine vs forged signatures.
- Implemented CNN architecture using VGG19 for feature extraction.
- Designed a web UI for uploading signatures and generating predictions.
  - Tech Stack: Python, Keras, CNN, VGG19, HTML, CSS, JavaScript