Data Engineer

Six months Course

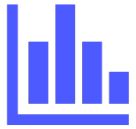
By

Nagendra Yadav

Role of a Data Engineer

Design, build, and maintain data pipelines

Ensure reliable data flow from multiple sources to storage systems



Support analytics, BI, and ML teams with clean, usable data

Optimize data infrastructure for scalability and performance

Role and Responsibility

Develop and schedule **ETL/ELT pipelines** using Python, SQL, and Spark

Ingest structured/unstructured data from APIs, files, or streams

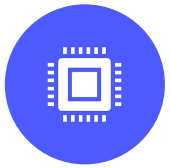
Design **data models** and manage **data warehouses** (BigQuery, Redshift)

Implement **workflow orchestration** with Airflow or Cloud Composer

Monitor data quality, data lineage, and pipeline failures

Manage storage on **cloud platforms** (GCP, AWS, Azure)

Collaborate with **data analysts, scientists, and product teams**



Apply **DevOps and CI/CD** for production-grade data solutions

Tools and software

VS code (IDE)

Python(programming)

Post-gre(SQL)

Spark(big data)

cloud

Github (Project repo.)

Nagendra Yadav

Problem Solving

DATA STRUCTURES & ALGORITHMS (DSA)

SQL QUERY CHALLENGES

DEBUGGING ETL FAILURES AND PERFORMANCE BOTTLENECKS

System Design

DESIGNING BATCH AND STREAMING PIPELINES

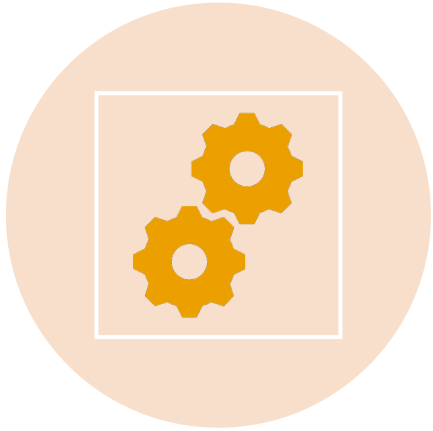
DATA WAREHOUSE/LAKEHOUSE ARCHITECTURE

SCALABLE AND FAULT-TOLERANT INGESTIONFRAMEWORKS

Core Engineering

PYTHON, PYSPARK FOR DATA PROCESSING

AIRFLOW FOR ORCHESTRATION



GIT, DOCKER, CI/CD TOOLS FOR DEPLOYMENT

Cloud & Big Data Tools

GCP (BigQuery, DataProc, Composer)

AWS (S3, Glue, Lambda, Redshift)

Spark, Hive, Kafka, HDFS

Business Intelligence Support

1

Build data models for BI tools (Power BI, Tableau)

2

Enable analytics and ML-ready datasets

3

Understand KPI/reporting requirements

Nagendra Yadav

Data Modeling & Governance

Star/Snowflake schema, SCDs

Data lineage, quality checks, documentation

Access control and audit logging (IAM roles)

Summary

Area	Tools/Skills
Languages	Python, SQL
Databases	PostgreSQL
Big Data	Hadoop, Hive, HDFS, Spark (PySpark)
Cloud	AWS/GCP/Azure
Streaming	Kafka, Spark Streaming
Orchestration	Airflow, Cloud Composer
CI/CD	Git, GitHub Actions, Docker
Dashboard	Tableau/Power BI

Tools link and Setup

- Python
- Postgres-SQL
- GitHub(for data engineer download and follow the instruction)
- Spark
- VS-code