Data Engineering Roadmap Checklist

Stage 1: Core Foundations

[ ] Learn SQL (SELECT, JOIN, GROUP BY, etc.)

[ ] Practice SQL with platforms like LeetCode SQL or Mode Analytics

[ ] Learn Python (focus on pandas, file I/O, functions)

[ ] Write simple ETL scripts in Python

[ ] Understand data modeling: normalized vs. denormalized

[ ] Learn star/snowflake schema

[ ] Learn Git basics (clone, commit, branch, merge)

Stage 2: Data Pipelines and Warehousing

[ ] Understand ETL vs. ELT workflows

[ ] Learn and use dbt for ELT pipelines

[ ] Learn Apache Airflow (or Prefect) for orchestration

[ ] Use a cloud data warehouse (BigQuery, Snowflake, or Redshift)

[ ] Learn about data lakes and file formats (Parquet, Avro)

Stage 3: Scalable Data Systems

[ ] Learn Apache Spark (PySpark)

[ ] Learn cloud services (AWS, GCP, or Azure: S3/GCS, EC2/Dataproc, etc.)

[ ] Learn Docker to containerize data jobs

[ ] Learn basic Terraform (infrastructure as code)

Stage 4: Advanced and Production-Grade

[ ] Set up CI/CD pipelines (GitHub Actions, GitLab CI)

[ ] Use dbt tests or Great Expectations for data quality

[ ] Monitor data pipelines (Prometheus, Grafana, Airflow logs)

[ ] Learn about streaming data (Kafka, Spark Streaming, Flink)

[ ] Understand data governance and security tools (Apache Atlas, Collibra)