Weekly Data Engineering Roadmap

Phase 1: Foundations (Weeks 1–8)

Goal: Build strong programming, SQL, and database fundamentals.

- **Week 1:** Python Basics
  - *Theory*: Variables, loops, functions, data structures.
  - *Practical*: Write scripts to clean CSV files.
- **Week 2:** Python for Data (pandas, NumPy)
  - *Theory*: DataFrames, indexing, transformations.
  - *Practical*: Load a dataset, clean missing values, calculate stats.
- **Week 3:** SQL Fundamentals
  - *Theory*: SELECT, WHERE, JOIN, GROUP BY.
  - *Practical*: Query a sample database (e.g., PostgreSQL).
- **Week 4:** Advanced SQL
  - *Theory*: Window functions, indexing, optimization.
  - *Practical*: Write queries to analyze sales trends.
- **Week 5:** Databases (Relational vs NoSQL)
  - *Theory*: Normalization, ACID properties, NoSQL basics.
  - *Practical*: Design a relational schema + insert/query data.
- **Week 6:** Linux & Shell Scripting
  - *Theory*: File system, permissions, cron jobs.
  - *Practical*: Automate a daily backup script.
- **Week 7:** Git & Version Control
  - *Theory*: Branching, merging, collaboration.
  - *Practical*: Push your ETL scripts to GitHub.
- **Week 8:** Mini Project

- *Practical*: Build a small ETL pipeline (CSV → Python transform → SQL database).

Phase 2: Core Data Engineering (Weeks 9–20)

Goal: Learn big data tools, orchestration, and cloud basics.

- **Week 9:** Introduction to Big Data
  - *Theory*: Hadoop ecosystem, distributed computing.
  - *Practical*: Run a simple MapReduce job.
- **Week 10:** Apache Spark Basics
  - *Theory*: RDDs, DataFrames, transformations.
  - *Practical*: Process a large dataset with Spark.
- **Week 11:** Spark Advanced
  - *Theory*: Spark SQL, MLlib basics.
  - *Practical*: Build a Spark job to aggregate logs.
- **Week 12:** Data Warehousing Concepts
  - *Theory*: OLAP vs OLTP, star schema, fact/dimension tables.
  - *Practical*: Design a warehouse schema for retail data.
- **Week 13:** Cloud Storage (AWS S3, Azure Blob)
  - *Theory*: Cloud storage principles, IAM roles.

    - *Practical*: Upload/download datasets from S3.
- **Week 14:** Workflow Orchestration (Airflow)
    - *Theory*: DAGs, scheduling, monitoring.
    - *Practical*: Automate ETL jobs with Airflow.
- **Week 15:** Data Modeling
    - *Theory*: Normalization vs denormalization, dimensional modeling.
    - *Practical*: Model a warehouse for a bank's transactions.
- **Week 16:** BI Tools (Power BI, Tableau)
    - *Theory*: Visualization principles, dashboards.
    - *Practical*: Build a dashboard from your warehouse.
- **Week 17–18:** Intermediate Project
    - *Practical*: End-to-end pipeline: ingest raw data → transform with Spark → store in warehouse → visualize in BI tool.
- **Week 19–20:** Cloud Basics (AWS/Azure)

- *Theory*: EC2, RDS, IAM, networking.
- *Practical*: Deploy your pipeline on AWS.

Phase 3: Advanced Engineering (Weeks 21–32)

Goal: Streaming, lakehouses, DevOps, advanced cloud.

- **Week 21:** Streaming Data Basics
    - *Theory*: Event-driven architecture, message queues.
    - *Practical*: Set up Kafka locally.
- **Week 22:** Kafka Advanced
    - *Theory*: Producers, consumers, topics, partitions.
    - *Practical*: Build a real-time log ingestion pipeline.
- **Week 23:** Data Lakes & Lakehouses
    - *Theory*: Delta Lake, Lakehouse architecture.
    - *Practical*: Create a data lake with S3 + Spark.
- **Week 24:** Cloud ETL Tools
    - *Theory*: AWS Glue, Azure Data Factory.
    - *Practical*: Build a cloud ETL job.
- **Week 25:** DevOps for Data Engineering
    - *Theory*: Docker, Kubernetes basics.
    - *Practical*: Containerize your ETL pipeline.
- **Week 26:** Infrastructure as Code
    - *Theory*: Terraform basics.
    - *Practical*: Deploy cloud resources with Terraform.
- **Week 27–28:** Advanced Project
    - *Practical*: Real-time analytics pipeline (Kafka → Spark → BI dashboard).
- **Week 29–30:** Security & Governance
    - *Theory*: POPIA compliance, IAM, encryption.
    - *Practical*: Secure your pipeline with IAM roles.
- **Week 31–32:** Cloud Specialization

- *Practical*: Choose AWS/Azure/GCP and build a full pipeline with their native tools.

Phase 4: Expert & Career Prep (Weeks 33–40)

Goal: Certifications, portfolio, and job readiness.

- **Week 33–34:** Certification Prep (AWS/Azure Data Engineer).
- **Week 35–36:** Portfolio Building
  - *Practical*: Document projects on GitHub + write case studies.
- **Week 37–38:** Mock Interviews & Problem Solving
  - *Theory*: Common data engineering interview questions.
  - *Practical*: Solve SQL + pipeline design challenges.
- **Week 39–40:** Capstone Project

- *Practical*: Build a **production-grade pipeline**: ingest streaming + batch data, store in warehouse, visualize, deploy on cloud.