Month 1: Foundations and Tools

Week 1: Introduction to Data Engineering and Ecosystem

- **Day 1: Roles and Responsibilities**
  - **What is a Data Engineer?**
    - Overview of the role within the data ecosystem.
    - Difference from Data Scientists and Analysts. (Optional)
    - Key deliverables: pipelines, infrastructure, and scalability.
  - **Core Responsibilities:**
    - Building and maintaining data pipelines.
    - Data integration, transformation, and storage.
    - Supporting downstream analytics and ML workflows.
  - **Scope in the Real World:**
    - Demand for Data Engineers in the industry.
    - Career paths and growth opportunities.
- **Day 2: Key Concepts**
  - **What is Data?**
    - Types: Structured, Semi-structured, Unstructured.
    - Formats: JSON, CSV, Parquet, Avro.
  - **Data Pipelines Overview:**
    - What are data pipelines and their role in the data ecosystem?
  - **ETL (Extract, Transform, Load):**
    - Why ETL is foundational to data workflows.
    - Example: Moving data from APIs to databases.
  - **Hands-On Introduction:**
    - Setting up Airbyte for simple data extraction.
    - Setting up Minio as a destination.

Week 2: ETL vs ELT

- **Day 1: ELT and Reverse ETL**
  - **What is ELT?**
    - Difference in process (Transformation post-load).
    - Use cases: Modern data platforms like Snowflake.
  - **Reverse ETL:**
    - Bringing transformed data back to operational systems.
    - Examples: Sending processed data back to Salesforce or other CRMs.
  - **Key Differences:**
    - Use case comparisons.
    - Cost and performance implications.
- **Day 2: Tools and Implementation**
  - **Hands-On ETL Tools:**
    - Extraction using Airbyte.

        - Loading to Minio using Iceberg.
        - Transformation using dbt:
            - Building simple transformations.
            - Writing SQL models.

Week 3: Data Infrastructure

- **Day 1: Databases**
    - **Relational Databases:**
        - Core concepts: Tables, indexes, primary keys, foreign keys.
        - Common tools: MySQL, PostgreSQL.
    - **NoSQL Databases:**
        - Core concepts: Key-value stores, document stores, graph databases.
        - Use cases: MongoDB, Cassandra.
- **Day 2: Data Lakes vs Warehouses**
    - **What is a Data Lake?**
        - Characteristics: Raw data storage, schema-on-read.
        - Use cases and tools: Hadoop, Iceberg.
    - **What is a Data Warehouse?**
        - Characteristics: Structured data, schema-on-write.
        - Use cases and tools: Redshift, Snowflake, BigQuery.
    - **Key Differences:**
        - Scalability, cost, and performance.

Week 4: Orchestration Tools

- **Day 1: Airflow Basics**
    - **What is Orchestration?**
        - Need for scheduling and automation.
        - Overview of Apache Airflow.
    - **Core Components:**
        - DAGs (Directed Acyclic Graphs).
        - Tasks and dependencies.
- **Day 2: Hands-On with Airflow**
    - Setting up Airflow locally.
    - Creating a simple DAG:
        - Tasks for data extraction, transformation, and loading.
        - Monitoring DAG runs.

Month 2: Data Modelling and Advanced Pipelines

Week 5: Data Modelling Concepts

- **Day 1: Introduction to Data Modelling**
  - **Banking Domain Overview:**
    - Types of data: Transactions, accounts, customers.
    - Business requirements for analytics and reporting.
  - **Star Schema vs Snowflake Schema:**
    - Differences, advantages, and trade-offs.
    - Examples for both schemas.
- **Day 2: Dimension Modelling**
  - **Fact Tables:**
    - Quantitative data (e.g., sales, transactions).
  - **Dimension Tables:**
    - Descriptive data (e.g., customer, product).
  - **ERD Tools:**
    - Creating models using Lucidchart or dbt.

Week 6: Building Data Models

- **Day 1: Hands-On Data Modelling**
  - Building a banking data model.
  - Identifying facts and dimensions.
- **Day 2: Validating Models**
  - Verifying relationships between tables.
  - Optimizing schema for performance.

Week 7: Advanced Data Pipelines

- **Day 1: Deep Dive into Extraction and Loading**
  - Advanced Airbyte usage:
    - Extracting data from multiple sources (APIs, files).
  - Loading to Minio with Iceberg:
    - Creating partitions and file optimization.
- **Day 2: Transformation**
  - Using dbt for advanced SQL-based transformations.
  - Using Pandas and PySpark for programmatic transformations.

Week 8: Reverse ETL

- **Day 1: Concepts and Tools**
  - Reverse ETL process and tools overview.
- **Day 2: Hands-On**
  - Example: Syncing processed data back to a CRM system.

Month 3: Data Quality, Testing, and Projects

Week 9: Data Quality and Testing

- **Day 1: Importance of Data Quality**
  - Common issues: Missing data, duplicates, inconsistencies.
  - Tools for quality checks.
- **Day 2: Unit Testing for ETL**
  - Writing tests for pipeline steps (e.g., transformation validation).

Week 10: Capstone Project Introduction

- **Day 1: Project Briefing**
  - Overview of telecom and banking projects.
  - Generating mock data.
- **Day 2: Setting Up Pipelines**
  - Starting ETL pipelines for the chosen project.

Week 11: Project Development

- **Day 1: Intermediate Steps**
  - Refining transformations and models.
- **Day 2: Integration and Orchestration**
  - Setting up final Airflow DAGs.

Week 12: Project Completion and Presentation

- **Day 1: Finalizing and Testing**
  - Data quality checks and pipeline validation.
- **Day 2: Presentations**
  - Students present their projects.
  - Feedback and suggestions for improvement.

Core Concepts Across Modules

1. **ETL/ELT:** Focus on real-world implementation and tool usage.
2. **Data Infrastructure:** Understanding databases, data lakes, and warehouses.
3. **Data Modelling:** Real-world schema design.
4. **Orchestration:** Automation with Airflow.
5. **Testing and Quality:** Building robust and reliable pipelines.

This detailed schedule ensures a balance between theory, hands-on practice, and project-based learning to build job-ready skills.