Cloud Data Engineering

Module 1: Data Acquisition

This module covers the core principles of acquiring, manipulating, and processing data from various sources, emphasizing automation and efficient web scraping techniques. You will learn how to set up your data engineering environment, explore Python for data manipulation, manage projects with version control, and gain hands on experience with web scraping using BeautifulSoup and Selenium. By the end of this module, you will be proficient in automating data acquisition and processing workflows.

1. Introduction to Data Engineering + Python Setup

Overview

- This section provides an introduction to data engineering, focusing on setting up the Python environment necessary for handling data.
- Learn the basic concepts of data engineering, the data lifecycle, and setting up tools for Python development.

2. Python for Data Engineering (Pandas) + Case Study

Overview:

- In this section, you will dive deeper into Python programming and data manipulation using the pandas library, an essential tool for data engineers.
- A case study will focus on cleaning, transforming, and analysing real-world datasets.

Case Study:

- Apply learned concepts to a real-world data problem by cleaning and analysing a dataset using pandas.

3. Version Control (Git + Python Project)

Overview:

- This module focuses on version control using Git and GitHub, essential for tracking changes in your code and collaborating with others.
- You will work on a Python project that integrates Git for managing changes and collaborating with peers.

Assignment:

- Python + Git assignment: Create a simple Python project, push it to GitHub, and collaborate on it using version control best practices.

4. Bash/Shell Scripting

Overview:

- Learn the basics of Bash and Shell scripting, which are essential for automating repetitive tasks and managing data pipelines.

- This section will equip you with the knowledge to write scripts for file manipulation, process automation, and task scheduling.

Assignment:

- Create a Bash script to automate a data acquisition task, such as downloading or processing files.

5. **Web Scraping with BeautifulSoup**

Overview:

- This section introduces the fundamentals of web scraping using BeautifulSoup. You will learn how to extract data from static websites and automate the process of retrieving data.

Assignment:

- Write a script using BeautifulSoup to scrape data from a static webpage, clean it, and save it to a structured format (CSV/JSON).

6. **Web Scraping with Selenium**

Overview:

- This module covers more advanced web scraping using Selenium, focusing on dynamic websites that rely on JavaScript for content generation. You will learn how to interact with websites, automate navigation, and extract complex data.

Assignment:

- Create a Selenium script to scrape data from a dynamic website (e.g., an ecommerce site), clean the data, and store it.

7. **Web Scraping Case Study**

Overview:

- This final section integrates the skills you've acquired in web scraping with BeautifulSoup and Selenium into a complete data acquisition pipeline. You will work on a case study involving both static and dynamic data extraction, processing, and analysis.

Final Assignment:

- Build a comprehensive web scraping solution to collect, process, and analysed data from a website (or multiple websites) using both BeautifulSoup and Selenium. Document your findings in a final report.

Module 2: Data Modelling

This module covers the fundamental principles of database design and management, emphasizing efficient data storage, retrieval, and optimization. You will learn data modelling techniques using SQL and gain a solid understanding of database normalization, relationships, and constraints. By the end of this module, you will be equipped to architect database structures that effectively address real-world data engineering challenges.

SQL with SQL Server:

1. **Fundamentals**
    - Introduction to SQL
    - Basic operations: SELECT, WHERE, ORDER BY, LIMIT
    - Implementing constraints for data integrity
2. **Data Definition and Manipulation**
    - Data Definition Language (DDL) for creating and altering database structures
    - Data Manipulation Language (DML) for modifying data within tables
3. **Advanced Querying Techniques**
    - Aggregation with GROUP BY
    - Set operations: UNION, INTERSECT, EXCEPT
    - Multidimensional analysis using CUBE and ROLLUP
4. **Joining Data**
    - Understanding and implementing various JOIN types:
        - Inner Join
        - Left (Outer) Join
        - Right (Outer) Join
        - Full (Outer) Join
5. **Performance and Structure**
    - Creating and managing INDEXES for query optimization
    - Writing and utilizing SUBQUERIES for complex data retrieval
    - Designing and implementing VIEWS for simplified data access
6. **Logical Operations and Automation**
    - Implementing conditional logic with CASE statements
    - Automating database actions using TRIGGERS
7. **Advanced SQL Concepts**
    - Structuring complex queries with Common Table Expressions (CTEs)
    - Performing advanced analytics using WINDOW FUNCTIONS
8. **Encapsulating Logic**
    - Creating and managing STORED PROCEDURES for reusable database operations

Module 3: Cloud Data Warehousing

This module explores cloud-based data warehousing with Snowflake, focusing on its architecture, scalability, and best practices for handling large datasets in the cloud. You will complete assignments to earn badges and work on multiple projects that help solidify your skills in Snowflake. By the end of this module, you will have hands-on experience with Snowflake and will be able to efficiently manage and query data in the cloud.

Snowflake

1. **Snowflake Overview**
    - Introduction to Snowflake architecture and key features.
    - Setting up your Snowflake environment for data warehousing.
2. **Badges Assignment**
    - Completion of badge tasks to build familiarity with Snowflake concepts and functions.
3. **Course Materials**
    - Follow along with the Snowflake Masterclass:
        - Snowflake Masterclass on Udemy
        - Complete the five sections of the course.
4. **Projects**
    - Apply your learning in **3 hands-on projects** to implement Snowflake data warehousing techniques.
    - Projects include:
        - Creating and optimizing data warehouses.
        - Data ingestion and ETL processes using Snowflake.
        - Advanced querying and performance tuning.

Module 4: Data Orchestration & Streaming

This module delves into managing data pipelines using Apache Airflow and real-time data streaming with Apache Kafka. You will learn how to build automated workflows and manage streaming data systems. Hands-on projects will allow you to gain practical experience in orchestrating complex data workflows and handling live data streams.

Airflow

1. **Introduction to Apache Airflow**
    - Learn the basics of workflow automation using Airflow.
    - Understand how to schedule, manage, and monitor complex data pipelines.
2. **Classes**
    - Two in-depth classes on Airflow covering workflow management, DAGs, operators, and tasks.
3. **Project**
    - Create a real-world project utilizing Airflow to orchestrate and automate data pipeline workflows.

Kafka

1. **Introduction to Apache Kafka**
    - Explore real-time data streaming concepts with Kafka.

- Learn how Kafka handles large-scale data streams and event-driven architectures.

2. **Classes**
   - Two classes on Kafka focusing on the basics of stream processing and building Kafka producers and consumers.
3. **Project**
   - Build a real-time data streaming project using Kafka, handling large data streams and demonstrating its use in distributed systems.

Module 5: Architecting AWS Data Engineering Projects

This module will provide an in-depth understanding of architecting and managing data engineering projects using various AWS services. You will work on **2–3 AWS projects** that incorporate data warehousing, ETL, and real-time data processing. By the end of the module, you will have developed the skills to design, implement, and scale AWS-based data engineering solutions.

AWS Redshift

- Learn how to build and manage cloud-based data warehouses using AWS Redshift.
- Topics covered include setting up clusters, data ingestion, querying, and performance optimization.

AWS S3

- Explore Amazon S3 for data storage and management.
- Learn how to efficiently store, retrieve, and manage data in S3 buckets for data engineering workflows.

AWS Glue & Athena

- Master data transformation and querying in AWS using Glue for ETL processes and Athena for SQL queries on data stored in S3.

AWS Lambda

- Automate data workflows with AWS Lambda, learning how to create serverless functions to trigger data processing and ETL jobs.

AWS EC2

- Set up and manage compute resources with EC2 instances to support data engineering tasks like running large-scale data processing jobs.

AWS RDS

- Use Amazon RDS to manage relational databases for structured data storage and querying.

AWS Kinesis

- Explore real-time data streaming with AWS Kinesis to handle large-scale event streams and build real-time applications.