AWS Data Engineering Program

Course Duration: 5 Days (8 Hours per Day)

Day 1: Data Ingestion and Transformation with AWS Glue

Objective: Build strong fundamentals in data ingestion, transformation, and ETL development using AWS Glue.
Modules:

1. **Introduction to AWS Data Engineering**
    - Overview of Data Engineering on AWS
    - Key AWS Services for Data Engineering (Glue, Redshift, Athena, DMS, Lake Formation, MWAA)
    - Data pipeline lifecycle and architecture patterns
2. **Introduction to AWS Glue**
    - Overview, Key Features, and Use Cases
    - AWS Glue Console Tour
    - Glue Data Catalog concepts and setup
    - Managing Metadata and Connections
3. **ETL and Data Transformation**
    - Creating and running ETL jobs in Glue
    - Data Transformation Concepts (Schema, Mapping, Joins)
    - Hands-on: Create a basic ETL job
4. **Working with Data Sources and Targets**
    - Connecting to AWS and external sources
    - Working with S3, RDS, Redshift as targets
    - Hands-on: Configure Glue job to load data into Redshift
5. **Optimization and Monitoring**
    - Job Scheduling and Orchestration (EventBridge, Triggers)
    - Logging, Debugging, and Performance Optimization
    - Cost optimization and scaling Glue jobs

Day 2: Advanced Glue, Crawlers, and Integration

Objective: Deep dive into automation, complex ETL flows, and integration with other AWS services.

1. **Advanced AWS Glue ETL Concepts**

- Complex Transformations
- Custom Scripts and Error Handling
- Partitioning and Parallel Processing

2. **AWS Glue Crawlers and Data Catalogs**
   - Creating and configuring crawlers
   - Schema detection and synchronization
   - Hands-on: Automate schema discovery
3. **Integration and Orchestration**
   - AWS Glue with Lambda, S3, and Step Functions
   - AWS Glue DataBrew for Data Quality and Profiling
   - Hands-on: Building an end-to-end pipeline using Glue and Lambda
4. **Best Practices**
   - Security and Access Control (IAM Roles, Secrets Manager)
   - Performance and Cost Optimization
   - Real-world Case Study: Enterprise ETL with Glue

Day 3: Data Lake Management with AWS Lake Formation

Objective: Learn to design, build, and secure data lakes using AWS Lake Formation.

1. **Introduction to Data Lakes and Modern Data Architecture**
   - What is a Data Lake?
   - Data Lake vs Data Warehouse
   - Modern Data Architecture using AWS services
2. **Getting Started with AWS Lake Formation**
   - Lake Formation Setup and Permissions Model
   - Ingesting and Registering Data in the Lake
   - Hands-on: Build a data lake from S3 using Lake Formation
3. **Data Governance and Cataloging**
   - Glue Catalog and Lake Formation Integration
   - Managing Permissions for Redshift, Athena, and EMR
   - Hands-on: Configure fine-grained access controls
4. **Automation and Blueprinting**

- Lake Formation Blueprints
- Automating Data Lake Creation
- Hands-on: Automate ingestion workflows

5. **Monitoring and Best Practices**
    - Auditing, Logging, and Security Controls
    - Cost and Performance Optimization
    - Modern data lake reference architecture

Day 4: Data Warehousing with Amazon Redshift

Objective: Learn how to build, optimize, and manage data warehouses in Redshift.

1. **Introduction to Amazon Redshift**
    - Architecture, Use Cases, and Console Overview
    - Data Distribution and Storage Concepts
2. **Data Ingestion and Loading**
    - Loading Data from S3 and Glue
    - Redshift Spectrum for Querying S3 Data
    - Hands-on: Load and query data in Redshift
3. **Querying and Transformation**
    - SQL Querying and Advanced Features (SUPER Data Type)
    - Optimization Techniques: Sort Keys, Dist Keys, Compression
    - Hands-on: Transform and analyze data
4. **Security and Monitoring**
    - IAM, Encryption, and Network Security
    - Monitoring and Troubleshooting using CloudWatch
    - Hands-on: Monitor query performance and logs
5. **Designing Analytics Solutions**
    - Data Warehouse Design Patterns
    - Integrating with QuickSight and Athena
    - Case Study: Analytics solution design

Day 5: Querying, Migration, and Workflow Automation

Objective: Bring together analytics (Athena), migration (DMS), and orchestration (MWAA) for a complete data lifecycle.

1. **Amazon Athena**
    - Overview and Architecture
    - Querying Data from S3 and Data Catalog
    - Hands-on: Query Parquet/CSV data in Athena
    - Integration with Glue Catalog and Lake Formation
2. **AWS Database Migration Service (DMS)**
    - Overview and Use Cases
    - Architecture and Components
    - Setting up Replication Instances, Source, and Target Endpoints
    - Hands-on: Migrate data from RDS to Redshift
    - Automating DMS with CLI/CloudFormation
3. **Managed Workflows for Apache Airflow (MWAA)**
    - Introduction to MWAA and Apache Airflow concepts
    - Orchestrating AWS Glue, Redshift, and DMS Jobs
    - DAG creation and dependency management
    - Hands-on: Build an MWAA DAG for end-to-end data pipeline
4. **Capstone Project**
    - Design and implement a complete data pipeline:
        - Ingest (Glue) → Store (Lake Formation) → Transform (Redshift) → Query (Athena) → Automate (MWAA)