Data Engineering in AWS

Course Duration: 20 hours

Course Syllabus:

1. Introduction to Database and Programming
2. Introduction to Data Engineering and AWS Services
3. Data Storage and Management
4. Data Integration and ETL with AWS Glue
5. Data Analytics and Querying
6. Real-Time Data Processing
7. Big Data Processing with AWS EMR
8. Security, Monitoring, and Optimization
9. Advanced Data Engineering with Azure Databricks
10. Security and Monitoring
11. Capstone Project

Tools & Technologies Covered:

1. SQL
2. Python
3. Amazon S3
4. RDS Database
5. Redshift
6. AWS Glue
7. Athena
8. QuickSight
9. Kinesis
10. Lambda
11. EMR
12. IAM
13. KMS
14. AWS CloudWatch

01 Introduction to Database and Programming

- SQL Basics
- Python – Basics, PySpark, Pandas, NumPy, Requests, Boto3

02 Introduction to Data Engineering and AWS Services

- Overview of data engineering and its role in modern businesses.
- Introduction to AWS for data engineering.
- Key services overview: S3, RDS, Redshift, Glue, EMR, Kinesis, Athena.
- **Hands-on Lab:**
  - Set up an AWS account and IAM roles for secure access.
  - Explore the AWS Management Console.

03 Data Storage and Management

- **Amazon S3:** Object storage fundamentals, lifecycle policies, and versioning.
- **AWS RDS:** Managed relational databases; PostgreSQL and MySQL basics.
- **Amazon Redshift:** Cloud data warehousing basics and architecture.
- **Hands-on Labs:**
  - Create S3 buckets and manage data (CSV, JSON).
  - Launch an RDS instance and connect to it.
  - Create a Redshift cluster and load data into it.

04 Data Integration and ETL with AWS Glue

- **AWS Glue:** ETL concepts, Glue catalog, and workflows.
- Data preparation and transformation using Glue Studio.
- Integration with S3 and Redshift.
- **Hands-on Labs:**
  - Set up Glue crawlers and catalog metadata.
  - Create and run ETL jobs to transform and load data from S3 to Redshift.

05 Data Analytics and Querying

- **Amazon Athena:** Serverless query service for analyzing S3 data.
- **AWS QuickSight:** BI tool for creating visualizations and dashboards.
- **Hands-on Labs:**
  - Query data stored in S3 using Athena.
  - Build a QuickSight dashboard with data from Athena or Redshift.

06 Real Time Data Processing

- **Amazon Kinesis:** Stream data ingestion and processing.
- **AWS Lambda:** Serverless compute for event-driven architectures.
- Use cases for real-time analytics and streaming.
- **Hands-on Labs:**
  - Create a Kinesis Data Stream and process data with Lambda.
  - Analyze real-time streaming data with Kinesis Analytics.

07 Big Data Processing with AWS EMR

- AWS EMR (Elastic MapReduce): Hadoop and Spark-based big data processing.
- Processing and analyzing large datasets using PySpark.
- **Hands-on Labs:**
  - Set up an EMR cluster.
  - Run a PySpark job to process large data files in S3.

08 Security, Monitoring, and Optimization

- **AWS Security:** IAM roles, S3 bucket policies, encryption, and KMS.
- **AWS CloudWatch:** Monitoring pipelines and setting alerts.
- Cost optimization: Strategies for reducing AWS costs.
- **Hands-on Labs:**
  - Implement bucket encryption and access control lists (ACLs).
  - Set up CloudWatch alarms and monitor data pipelines.

09 Capstone Project

T**itle:** Sales Reporting Pipeline in AWS

Objective: Build a complete pipeline to ingest, process, and analyze sales data using AWS services.

- **Data Ingestion:** Use AWS Glue to ingest sales data from Amazon S3 and customer data from Amazon RDS.
- **Data Storage:** Store raw data in Amazon S3 for scalable storage.

- **Data Transformation:** Process and aggregate data using AWS Glue (Spark-based).

- **Data Loading:** Load the transformed data into Amazon Redshift.

- **Reporting & Visualization:** Use Amazon QuickSight to create dashboards and visualizations.