



AWS GLUE TRAINING

**DURATION OF
TRAINING: 24-30HRS.
BATCH TYPE:
WEEKDAYS/WEEKENDS
MODE OF TRAINING:
CLASSROOM/ONLINE/
CORPORATE TRAINING**

Module 1: Introduction to AWS Glue

- What is AWS Glue? Use Cases
- Glue Architecture & Components Overview
- Comparison with other ETL services (e.g., Data Pipeline, EMR)
- Glue vs Lambda vs Step Functions for ETL

Hands-on:

- Explore AWS Glue Console and understand regions & pricing

Module 2: Data Catalog

- Glue Data Catalog Overview
- Crawlers and their Configuration
- Classifiers and Custom Classifiers
- Table Metadata and Partitions
- Glue Data Catalog Integration with Athena, Redshift, EMR

Hands-on:

- Create Glue Crawlers to catalog S3 data
- Query data using Athena via Glue Catalog

Module 3: Glue Jobs – Python & Spark ETL

- Glue Job Types (Spark, Python Shell, Ray)
- Glue DynamicFrame vs Spark DataFrame
- Connection to RDS, Redshift, S3
- Reading, Writing & Transforming data
- Bookmarking and Job Parameters
- Glue Job Bookmarks and Checkpoints

Hands-on:

- Create a Glue Spark Job to transform JSON to Parquet
- Load data from S3 to Redshift using Glue Job

Module 4: Glue Studio (No-Code/Low-Code Development)

- Introduction to Glue Studio Interface
- Visual Job Creation
- Adding and Testing Transforms
- Monitoring Job Execution & Logs

Hands-on:

- Build an ETL pipeline using Glue Studio without writing code

Module 5: Job Orchestration & Triggers

- Creating Triggers (On-Demand, Scheduled, Conditional)
- Job Dependencies and Event Triggers
- Integration with Step Functions

Hands-on:

- Chain multiple Glue jobs using conditional triggers
- Set up a Step Function to run jobs in sequence

Module 6: Glue DataBrew (Data Wrangling without Code)

- What is AWS Glue DataBrew?
- Creating Projects and Recipes
- Visualizing and Cleaning Data
- Exporting Cleaned Data to S3/Redshift

Hands-on:

- Clean CSV file using Glue DataBrew (null removal, column cleanup)

Module 7: Glue Security & Access Control

- IAM Roles and Policies for Glue
- Encrypting data at rest and in transit
- VPC Endpoints and Glue Networking
- Data Classification for DLP

Hands-on:

- Assign IAM role for S3 access to Glue
- Configure KMS encryption for Glue jobs

Module 8: Monitoring, Logging & Optimization

- Job Metrics and Logs (CloudWatch)
- Debugging Errors in Jobs
- Optimizing Job Execution (Worker Types, DPUs, Retry Policies)
- Cost Optimization Strategies

Hands-on:

- Enable CloudWatch logs and alerts for failed jobs
- Tune Glue job for large file batch processing

Module 9: Advanced Use Cases & Integration

- Partitioning Strategies and Incremental Loads
- Handling Semi-structured Data (JSON, XML)
- Glue with AWS Lake Formation
- Glue with S3 Event Notification
- Glue + Redshift + Athena Pipelines

Capstone Project

Project Scenario:

Design a serverless ETL pipeline using AWS Glue that:

- Crawls and catalogs raw CSV logs from S3
- Cleans and transforms the data into Parquet
- Loads it into Redshift for reporting
- Triggers the ETL job daily via schedule
- Includes monitoring, logging, and failure notifications

Tools & Technologies Covered

- AWS Glue (Jobs, Catalog, Studio, DataBrew)
- S3, Redshift, Athena, RDS
- AWS IAM, CloudWatch, Step Functions
- Python (PySpark in Glue Jobs)

Optional Add-Ons

- Glue + Lake Formation Integration
- Glue + DMS for CDC (Change Data Capture)
- Build ETL using Glue + Apache Hudi or Iceberg

Phone : +91 8055223360

Email : training@radicaltechnologies.co.in

Website : www.radicaltechnologies.co.in

PUNE | BANGALORE | KERALA | UK

PUNE - (AUNDH, KHARADI & HINJEWADI)

KERALA - (CALICUT, KOCHI & TRIVANDRUM)