

AWS Data Engineering

10-Week / 50-Topic Detailed Professional Syllabus

Course Overview

Course Focus	PySpark + AWS + Databricks + Streaming + DevOps + AI/Claude Code
Structure	10 weeks, 50 weekday topics, weekend masterclasses, and 5 major projects
Learning Style	Concepts + live coding + architecture + troubleshooting + project implementation

Detailed Weekly Curriculum

Week 1 – PySpark & Databricks Foundations

- 1. PySpark & Databricks Introduction
- 2. Databricks Functions (date, string, numeric)
- 3. Window Functions (rank, lag, lead)
- 4. Data Cleaning with Regex
- 5. Spark-submit Internals

Weekend Masterclass: Claude Code + Prompt Engineering
Hands-on session exploring AI-assisted coding, generating/reviewing PySpark, debugging, and prompt patterns for data engineering.

Week 2 – Advanced Spark Data Processing

6. Spark JSON Processing
7. Spark JDBC (Oracle, MySQL, PostgreSQL via RDS)
8. XML, CSV & Pandas Data Processing
9. Parquet & ORC
10. Delta Lake vs Iceberg

Weekend Masterclass: File Format & Performance Lab

Deep dive into columnar storage, execution plans, partition pruning, and optimizing read/write performance in Spark.

Week 3 – AWS Storage & Integration

11. AWS Introduction (Accounts, Regions, IAM basics)
12. S3 Buckets, Policies & Data Lake Design
13. AWS Glue Introduction & Internals
14. Glue Jobs + Databricks Integration
15. Parameterization & Glue Data Catalog

Weekend Masterclass: Metadata-Driven Glue Pipelines

Design and parameterize reusable Glue pipelines driven by configuration metadata.

Week 4 – AWS Glue Advanced & End-to-End Project

16. IAM Users, Roles, Policies
17. Secrets Management (AWS Secrets Manager vs Databricks Secrets)
18. Databricks + S3 Integration (IAM roles, instance profiles)
19. Glue Advanced Concepts
 - Crawlers & schema evolution
 - Partitioning strategies
 - Job bookmarks & incremental loads
 - Error handling, retries & best practices for large datasets

20. End-to-End Batch Pipeline (S3 → Glue → Databricks → Delta)

Weekend Masterclass: Glue Metadata-Driven Pipelines

Advanced implementations of metadata-driven architectures utilizing AWS Glue components.

Week 5 – Delta Lake & Databricks Optimization

21. Databricks Utilities (dbutils.fs, secrets, widgets)

22. Delta Time Travel & Change Data Feed

23. Medallion Architecture on AWS (Bronze/Silver/Gold)

24. Declarative Pipelines with Databricks

25. Databricks Optimization (OPTIMIZE, Z-Ordering, Photon)

Weekend Masterclass: Delta Sharing & Lakehouse Federation

Secure data sharing, federated query use cases, and governance considerations across lakehouses.

Week 6 – AWS Orchestration & Databricks Deployment

26. AWS Lambda for ETL triggers

27. AWS Step Functions for orchestration

28. Advanced Step Functions (error handling, retries)

29. Databricks Asset Bundles (DABs)

30. Unity Catalog Advanced (governance, row filters, masks)

Weekend Masterclass: CI/CD with GitHub Actions + DABs

Implementing Git-based development, automated testing, and multi-environment deployment using DABs.

Week 7 – Streaming on AWS

- 31. Spark Structured Streaming Introduction
- 32. Kafka with Spark (MSK)
- 33. NiFi + Kafka + Spark Streaming
- 34. AWS Kinesis & EventBridge Streaming Project
- 35. Auto Loader + Streaming Internals

Weekend Masterclass: Real-Time AWS Streaming Project

End-to-end streaming architecture with MSK/Kinesis into Databricks with checkpointing and state management.

Week 8 – Airflow, Testing & Performance

- 36. Apache Airflow Introduction (on EC2/EKS)
- 37. Airflow Advanced Orchestration (Glue + Databricks)
- 38. Testing & Observability (CloudWatch + Databricks Monitor)
- 39. Performance Tuning Deep Dive (joins, skew, AQE)
- 40. Spark RDD Internals

Weekend Masterclass: Production Troubleshooting Workshop

Diagnose slow jobs, trace pipeline failures, analyze logs, and prioritize performance fixes.

Week 9 – AWS Analytics & BI

- 41. Redshift Core Architecture
- 42. Redshift vs Snowflake vs Databricks Lakehouse
- 43. Real-Time Analytics with Kinesis + Lambda
- 44. Semantic Models with QuickSight
- 45. Preparing AI-Ready Analytics Data in AWS

Weekend Masterclass: AWS vs Databricks Architecture Workshop

Comparative analysis of AWS-native tools vs Databricks Lakehouse, evaluating cost, migration patterns, and architecture choices.

Week 10 – AI-Assisted Engineering & Interview Prep

- 46. Claude Code for Data Engineers
- 47. Prompt Engineering for ETL/SQL
- 48. GitHub Copilot for Productivity
- 49. GitHub Actions CI/CD with Databricks + AWS
- 50. Interview Tips & Resume Preparation

Weekend Masterclass: Mock AWS Data Engineer Interview Clinic

Practice technical scenarios, architecture explanations, and mock interviews to prepare for AWS Data Engineering roles.

Hands-On Projects

Project 1: Glue + Databricks Batch Pipeline

Pipeline: S3 → Glue → Databricks → Delta

Project 2: Metadata-Driven Glue Framework

Pipeline: Config → Glue → Databricks → Logging

Project 3: AWS Databricks Lakehouse

Pipeline: S3 → Auto Loader → Delta → Unity Catalog

Project 4: AWS Real-Time Streaming

Pipeline: Kafka/Kinesis → Databricks → Delta → Gold

Project 5: CI/CD & Production Deployment

Pipeline: GitHub Actions → DABs → AWS Environments

Technology Stack Covered

Core Engineering	PySpark, Spark SQL, JDBC, JSON, XML, Parquet, ORC, Delta Lake, Iceberg
AWS Storage	S3, Glue Data Catalog, RDS
AWS Integration	Glue, Lambda, Step Functions, IAM, Secrets Manager
AWS Streaming	Kinesis, MSK (Kafka), EventBridge, NiFi, Structured Streaming
Databricks	Databricks, Delta Lake, Unity Catalog, Lakeflow, DABs
Governance & Security	IAM, Secrets Manager, Unity Catalog
Orchestration & Ops	Airflow, Step Functions, Databricks Workflows
DevOps & AI	GitHub, GitHub Actions, Claude Code, Copilot, Prompt Engineering