

## Big Data Engineering – Advanced (2 Months / 8 Weeks)

**Duration:** 2 Months (Mon–Fri, ~80–90 Hours)

**Mode:** Live Online / Classroom

**Tools & Technologies:** Kafka, Databricks, Airflow, AWS SNS, Power BI/Tableau

### Syllabus

#### Week 1: Kafka Basics

- Kafka architecture (topics, partitions, brokers)
- Producers & Consumers
- Hands-on: Ingest sample data into Kafka topics

#### Week 2: Kafka Integration with Spark

- Kafka → Spark Structured Streaming ingestion
- Running streaming ETL jobs in PySpark
- Hands-on: Real-time pipeline from Kafka → Spark

#### Week 3: Databricks for Big Data Engineering

- Databricks architecture (clusters, notebooks, jobs)
- Running PySpark jobs on Databricks
- Using Delta Lake in Databricks (schema enforcement, time travel)
- Hands-on: Batch ETL pipeline in Databricks

#### Week 4: Advanced Databricks Use Cases

- Integrating Kafka streams into Databricks
- Optimizing Delta tables (merge, upserts, deletes)
- Job scheduling and monitoring in Databricks
- Mini Project: Near real-time ETL with Databricks

#### Week 5: Airflow Fundamentals

- Airflow architecture (scheduler, webserver, workers)
- Writing first DAGs for Spark jobs
- Operators, tasks, and dependencies
- Hands-on: Simple batch workflow in Airflow

### **Week 6: Airflow + Databricks Orchestration**

- Orchestrating Databricks jobs with Airflow operators
- Error handling, retries, and SLAs
- CI/CD best practices for Airflow workflows
- Case Study: Multi-stage ETL orchestrated by Airflow

### **Week 7: AWS SNS Integration**

- AWS SNS basics (topics, subscriptions)
- Sending notifications for pipeline success/failure
- Integrating SNS with Airflow for alerts
- Hands-on: Alert-enabled pipeline

### **Week 8: Capstone Project & Mock Interview 3**

- End-to-End Big Data Project:
- Ingestion: Kafka → S3
- Processing: PySpark on Databricks
- Storage: Delta Lake
- Orchestration: Airflow
- Notifications: AWS SNS
- Reporting: Power BI / Tableau
- Final Capstone Presentation
- Resume prep & GitHub publishing
- Mock Interview 3 (End-to-End Big Data Pipeline)

### **Learning Outcomes**

- Build real-time ingestion pipelines with Kafka
- Master Databricks for batch & near real-time ETL
- Implement Delta Lake optimizations (merge, schema enforcement, time travel)
- Orchestrate pipelines with Airflow + Databricks integration
- Integrate AWS SNS alerts into Big Data workflows
- Deliver end-to-end Big Data projects with reporting