

Big Data Engineering – Intermediate (2 Months / 8 Weeks)

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

Mode: Live Online / Classroom

Tools & Technologies: Apache Spark, PySpark, Databricks, Delta Lake, AWS S3

Syllabus

Week 1: Spark Fundamentals

- Apache Spark architecture (RDD, DAG, Catalyst optimizer)
- Spark installation / cluster basics
- RDD operations (map, filter, reduce, flatMap)
- Hands-on: WordCount with RDDs

Week 2: DataFrames in Spark

- DataFrames vs RDDs
- Creating DataFrames from CSV/JSON
- Column operations & filtering
- Hands-on: DataFrame transformations

Week 3: SparkSQL Basics

- Running SQL queries inside Spark
- SELECT, WHERE, Joins, GroupBy
- Optimizations with Catalyst & Tungsten
- Hands-on: SparkSQL on sales dataset

Week 4: Advanced SparkSQL

- Window functions (ROW_NUMBER, RANK)
- Aggregations with multiple columns
- Case Study: Customer behavior analysis

Week 5: PySpark for ETL

- PySpark DataFrame API
- Data ingestion, transformation, writing back
- Cleaning & deduplication at scale
- Hands-on: PySpark ETL project

Week 6: Databricks Introduction

- Databricks clusters & notebooks
- Writing jobs in Databricks
- Managing jobs & versioning notebooks
- Mini Project: ETL pipeline on Databricks

Week 7: Delta Lake Concepts

- Why Delta Lake?
- Time Travel & Schema Enforcement
- Implementing upserts & deletes with Delta
- Hands-on: Delta Lake integration with PySpark

Week 8: AWS S3 Integration & Wrap-Up

- AWS S3 as a Data Lake
- Reading/writing Spark data to S3
- Securing access with IAM
- Module recap + Mock Interview 2 (Spark + Databricks + S3)

Learning Outcomes

- Work with Spark RDDs, DataFrames, and SQL
- Build PySpark ETL pipelines
- Use Databricks for collaborative data engineering
- Integrate Spark with AWS S3 & Delta Lake