# Databricks Data Engineering Course

## 2-Month Program Outline

## Objective

Transform learners into **Data Engineers** using Databricks capable of:

- Processing large-scale data
- Building ETL pipelines
- Working with Spark (PySpark)
- Creating data lakes & analytics systems

## Program Duration

**8 Weeks (2 Months)**

### Structure

- **2 Days** → Learning
- **2 Days** → Hands-on Labs
- **1 Day** → Demo + Knowledge Sharing

## PHASE 1: Data Engineering Fundamentals (Week 1)

### Topics

- What is Data Engineering
- ETL vs ELT
- Data Warehouses vs Data Lakes
- Intro to Big Data concepts

### Tools & Labs

- **Tools:** Databricks, Python basics (for data)
- **Hands-on:** Setup Databricks workspace, Run first notebook, Load sample dataset

**Outcome:** Understand the data ecosystem and navigate Databricks.

## PHASE 2: PySpark Basics (Week 2–3)

### Topics

- Intro to Apache Spark
- RDD vs DataFrame
- PySpark basics
- Transformations & actions
- Data cleaning

### Tools & Labs

- **Tools:** PySpark, Databricks notebooks
- **Hands-on:** Read CSV/JSON data, Filter & transform data, Aggregate datasets

**Outcome:** Process large datasets and write PySpark jobs.

## PHASE 3: Data Processing & ETL Pipelines (Week 4)

### Topics

- ETL pipeline design
- Batch processing
- Data validation
- Schema handling

### Labs

- **Hands-on:** Build ETL pipeline, Clean raw data to structured format, Save processed data

**Outcome:** Build real-world data pipelines.

## PHASE 4: Delta Lake + Advanced Storage (Week 5)

### Topics

- Delta Lake
- ACID transactions
- Time travel
- Data versioning

### Labs

- **Hands-on:** Create Delta tables, Perform updates & deletes, Query historical data

**Outcome:** Manage reliable data lakes and handle large-scale storage.

## PHASE 5: Streaming Data (Week 6)

### Topics

- Streaming vs batch processing
- Structured Streaming
- Real-time data pipelines

### Labs

- **Hands-on:** Stream data (logs/events), Process real-time data, Store streaming output

**Outcome:** Build real-time data systems.

## PHASE 6: Data Integration + Cloud (Week 7)

### Topics

- Integrating with cloud (AWS / Azure)
- Data ingestion from APIs / DBs
- Scheduling pipelines

### Tools & Labs

- **Tools:** Amazon Web Services, Azure Databricks
- **Hands-on:** Connect Databricks to cloud storage (S3), Load data from external sources, Automate pipeline execution

**Outcome:** Build cloud-based data systems.

## PHASE 7: Optimization + Production (Week 8)

### Topics

- Performance tuning (Spark optimization)

– Partitioning & caching
– Job scheduling
– Monitoring pipelines

### Labs

– **Hands-on:** Optimize queries, Schedule jobs, Monitor pipeline performance

**Outcome:** Deliver production-ready data pipelines.

## FINAL PROJECT (End of Course)

*Choose ONE of the following options to build end-to-end:*

1. **Data Pipeline System**: Ingest raw data, Transform using PySpark, Store in Delta Lake
2. **Real-Time Analytics System**: Stream data, Process events, Dashboard output
3. **Data Warehouse Pipeline**: ETL pipeline, Structured data storage, Query system

## Recommended Tech Stack

- **Platform**: Databricks
- **Processing**: Apache Spark, PySpark
- **Storage**: Delta Lake
- **Cloud**: Amazon Web Services / Azure

## High-Value POC Categorization

- **Beginner**: Data cleaning scripts, CSV processing
- **Intermediate**: ETL pipelines, Data transformations
- **Advanced**: Streaming pipelines, Optimized Spark jobs

## Weekly Execution Plan

| Day | Activity |
|---|---|
| Monday | Learning |
| Tuesday | Learning |
| Wednesday | Hands-on Labs |
| Thursday | Hands-on Labs |
| Friday | Demo / Knowledge Sharing |

## Team Setup & Role Rotation

- Data Engineer
- PySpark Developer
- Cloud Integrator
- Pipeline Manager

## Evaluation Metrics

✓ Pipeline correctness

✓ Data accuracy

✓ Performance optimization

✓ Real-world usability

## Outcome After 2 Months

– Build ETL pipelines

– Process big data using Spark

– Work with Databricks

– Create production-ready data systems

## How This Fits Your Full Learning Stack

*From your previous roadmaps:*

→ **Frontend** → UI

→ **Backend (Java/Python)** → APIs

→ **AWS** → Cloud

→ **Databricks** → Data Engineering

**Final Result:** **Full Stack + AI + Cloud + Data Engineer**