🧩**Data Engineering program– From Python to PySpark on Databricks**
🎺**Start Date:** Nov 23 | 🎣**Duration:** 16 Weeks | **Platform:** Databricks + GitHub + Cloud

🚀Program Structure Overview

Phase	Timeline	Focus Area	Key Outcomes
Phase 1: Python on Databricks	Week 1 – 4	Python for Data Engineering	Write & run Python on Databricks, explore data handling, file I/O, and functions
Phase 2: SQL Mastery	Week 5 – 10	Basic + Advanced SQL	Master SQL joins, CTEs, subqueries, and window functions using company dataset
Phase 3: PySpark & Capstone	Week 11 – 16	PySpark + Cloud + Databricks	Build end-to-end ETL pipelines and deploy a Capstone project on Databricks

Weekly Breakdown

Week	Theme	Core Skills & Tools
1	Intro to Databricks + Python Basics	Variables, Data types, DBFS navigation, dbutils
2	Control Flow & Functions	Loops, Functions, Logging, Code Reuse
3	Collections in Python	Lists, Dictionaries, Sets, Comprehensions, Data Simulation
4	File Handling + Pandas	Read/Write CSVs, Data cleaning in Databricks
5	SQL on Databricks	CREATE/INSERT, SELECT, WHERE, ORDER BY
6	Aggregations & Filters	GROUP BY, HAVING, CASE WHEN, DATE ops
7	Joins Deep Dive	INNER, LEFT, RIGHT, FULL, SELF Joins
8	Subqueries & CTEs	Nested queries, WITH clause, correlated queries
9	Window Functions	RANK, DENSE_RANK, LEAD, LAG, SUM OVER
10	SQL Optimization	Indexing, Partitioning, Query refactoring
11	PySpark Fundamentals	SparkSession, Transformations, Actions
12	DataFrame Operations	Filtering, GroupBy, Aggregations, Null handling
13	Advanced PySpark	UDFs, Window functions, Broadcast joins, Performance tuning
14	Databricks Delta + Cloud	Delta tables, Versioning, Cloud storage integration
15	Capstone Project – Design	Architecture, Ingestion, Logging, Version Control
16	Capstone Project – Delivery	CI/CD (GitHub Actions), Testing, Documentation

GitHub Workflow Timeline

Milestone	Activity	Week
Repo Setup	Create folders: `/Python`, `/SQL`, `/PySpark`, `/Capstone`	Week 1
Weekly Commits	Push notebooks (.dbc/.ipynb) to GitHub	Ongoing
CI/CD Pipeline	Add GitHub Actions for ETL validation	Week 10
Capstone Delivery	Push full project with documentation	Week 16

By the End of the program You Will...

Write production-ready Python & SQL on Databricks
Build scalable PySpark ETL pipelines
Integrate with Cloud data storage
Collaborate with GitHub version control
Deploy an end-to-end Data Engineering Capstone Project