

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

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

Mode: Live Online / Classroom

Tools & Technologies: SQL, Python, Pandas, NumPy, Hive (Intro), PySpark (Intro)

Syllabus

Week 1: Big Data Foundations

- What is Big Data? Batch vs Streaming processing
- Hadoop ecosystem overview (HDFS, YARN, Hive, Spark)
- Why Spark over MapReduce?
- Industry use cases of Big Data

Week 2: SQL Basics (Part 1)

- RDBMS concepts refresher
- SQL DDL, DML (CREATE, INSERT, UPDATE, DELETE)
- SELECT statements & WHERE filters
- ORDER BY, LIMIT usage
- Hands-on: Run SQL queries on sample datasets

Week 3: SQL Basics (Part 2)

- Aggregations: SUM, COUNT, AVG, MIN, MAX
- GROUP BY, HAVING
- Combining filters with aggregations
- Case Study: Retail/Banking dataset analysis

Week 4: SQL Joins

- INNER, LEFT, RIGHT, FULL Joins
- Hands-on: Joining Orders + Customers dataset
- Joins with multiple tables
- Business case study queries

Week 5: Introduction to Hive

- Hive architecture & components (Metastore, HDFS)
- Difference between SQL & HiveQL

- Creating & loading Hive tables
- Hands-on: Query structured data in Hive

Week 6: Hive Queries & Integrations

- Hive DDL/DML commands
- Partitioning & Bucketing basics
- Performance considerations in Hive
- Mini Project: Sales dataset analysis in Hive

Week 7: Python for Data Engineering

- Python syntax, loops, functions, file handling
- Pandas & NumPy basics for analysis
- Hands-on: Cleaning CSV/JSON with Pandas

Week 8: Intro to PySpark & Wrap-Up

- What is PySpark? Scaling Python to Big Data
- Creating Spark DataFrames
- Simple transformations (select, filter, withColumn)
- Module recap & review
- Mock Interview 1 (SQL + Hive + Python basics)

Learning Outcomes

- Understand Big Data ecosystem and Hadoop basics
- Write SQL queries up to Joins
- Run SQL-like queries in Hive on large datasets
- Use Python, Pandas & NumPy for transformations
- Introduction to PySpark for distributed ETL