PySpark Learning Curriculum

Level 1: Introduction to Big Data

- What is Big Data? (Volume, Velocity, Variety, Veracity, Value)

- Structured vs. Unstructured Data

- Why Big Data Matters? Real-world Use Cases

- Traditional Systems vs Big Data Systems

- Big Data Ecosystem Overview (Hadoop, Spark, Hive, Kafka)

Level 2: Big Data Technologies & Architecture

- Hadoop Ecosystem: HDFS, MapReduce, YARN, Hive, Pig, HBase

- Apache Spark Overview: Core, SQL, Streaming, MLlib, GraphX

- Spark vs Hadoop

Level 3: Data Ingestion and ETL

- Batch vs Real-time Processing

- Apache Sqoop: RDBMS to Hadoop

- Apache Flume: Real-time Logs to HDFS

- Apache Kafka: Distributed Streaming

- Apache NiFi: Dataflow Automation

Level 4: Big Data Storage & NoSQL

- NoSQL Overview: Key-value, Document, Columnar, Graph DBs

- Examples: Redis, MongoDB, HBase, Cassandra, Neo4j

- Partitioning, Sharding, CAP Theorem

Level 5: Big Data Analytics & Processing

- Batch Processing: Spark, Hive, Pig

- Real-time Processing: Spark Streaming, Kafka Streams, Flink

- Data Aggregation, Joins, Filtering, SQL in Spark

Level 6: Big Data with Machine Learning

- MLlib in Spark: Pipelines, Transformers
- Training & Evaluating ML Models at Scale
- Streaming ML with Spark Streaming
- PySpark ML Integration

Level 7: Big Data on Cloud

- AWS: EMR, S3, Kinesis, Redshift
- Azure: HDInsight, Databricks
- GCP: BigQuery, Dataflow, Dataproc
- Serverless Big Data Tools

Level 8: Data Governance & Security

- Data Privacy (GDPR, HIPAA)
- Authentication & Authorization (Kerberos, Ranger)
- Data Encryption, Masking & Compliance