PySpark Learning Curriculum

Level 1: Introduction to Apache Spark & PySpark

- What is Apache Spark?
- Spark vs Hadoop MapReduce
- Spark Core Concepts: RDD, DAG, Lazy Evaluation
- Spark Ecosystem Overview
- What is PySpark? Installation & Setup

Level 2: PySpark Core - RDD API

- Creating RDDs: from files, collections
- RDD Operations: map, filter, flatMap, union, distinct
- Actions: collect, count, take, reduce, foreach
- Caching and Persistence
- Partitioning and Repartitioning

Level 3: PySpark SQL & DataFrames

- Creating DataFrames from CSV, JSON, RDDs
- DataFrame Operations: select, filter, groupBy, agg, orderBy, join
- Working with nulls and missing data
- Column manipulation and UDFs
- SQL Queries using spark.sql()

Level 4: Data Preprocessing & ETL with PySpark

- Reading and Writing Files: CSV, JSON, Parquet, Avro
- Data Cleaning and String manipulations
- Handling Nested Data Types
- Partitioning and Bucketing

Level 5: PySpark MLlib - Machine Learning

- MLlib Basics: Transformers, Estimators, Pipelines
- Feature Engineering: VectorAssembler, StringIndexer, OneHotEncoder

PySpark Learning Curriculum

- Algorithms: Logistic Regression, Random Forest, KMeans, Linear Regression
- Model Evaluation and Cross-validation

Level 6: PySpark Streaming

- Structured Streaming basics
- Reading data from Kafka, sockets, and files
- Window operations, aggregations, watermarking
- Output sinks: console, file, Kafka

Level 7: PySpark on Cloud & Performance Tuning

- Running PySpark on AWS, GCP, Azure
- Using Databricks
- Performance Tuning: partitioning, broadcast joins, caching

Level 8: PySpark Best Practices

- Writing modular PySpark code
- Logging and error handling
- Unit testing PySpark jobs
- PySpark with Airflow