PYSpark

1. Introduction

- *Overview of PySpark*
- *Importance of PySpark in big data processing*
- *Key features of PySpark*

2. Getting Started

- *Installation*
    - *System requirements*
    - *Installing Java*
    - *Downloading and setting up Apache Spark*
    - *Installing PySpark via pip*
- *Setting up the environment*
    - *Configuring environment variables*
    - *Verifying the installation*

3. Spark Basics

- *Spark Architecture*
    - *Spark components (Driver, Executors, Cluster Manager)*
    - *Spark ecosystem (HDFS, YARN, Mesos, etc.)*
- *Spark Concepts*
    - *Resilient Distributed Datasets (RDDs)*
    - *DataFrames*
    - *Datasets*
    - *Spark SQL*

4. Initializing Spark

- *Using SparkSession*
- *Configuration options*
- *Running Spark in different modes (local, standalone, cluster)*

5. Working with RDDs

- *Creating RDDs*
    - *From existing collections*
    - *From external datasets*

- *RDD Operations*
    - *Transformations (map, filter, flatMap, etc.)*
    - *Actions (collect, reduce, count, etc.)*
- *Persistence and Caching*
- *Key-Value Pair RDDs*

6. Working with DataFrames

- *Creating DataFrames*
    - *From RDDs*
    - *From structured data files (CSV, JSON, Parquet, etc.)*
- *DataFrame Operations*
    - *Selecting columns*
    - *Filtering rows*
    - *Grouping and aggregation*
    - *Joining DataFrames*
- *Working with SQL in PySpark*
    - *Registering DataFrames as SQL tables*
    - *Executing SQL queries*

7. Working with Datasets

- *Overview of Datasets*
- *Creating Datasets*
- *Transformations and Actions on Datasets*

8. Advanced Data Processing

- *Working with complex data types*
- *User-defined functions (UDFs)*
- *Window functions*
- *Pivot and Unpivot operations*

9. Machine Learning with PySpark MLlib

- Overview of MLlib
- Data preprocessing
- Feature engineering
- Building and evaluating models
- Model persistence and deployment

10. Graph Processing with GraphX

- Overview of GraphX
- Creating graph data structures
- Graph algorithms and operations

11. Structured Streaming

- Overview of Structured Streaming
- Creating streaming DataFrames
- Streaming transformations and actions
- Managing streaming queries

12. Performance Tuning

- Understanding Spark jobs and stages
- Optimizing transformations and actions
- Memory management
- Configuring Spark for performance

13. Deployment

- Running PySpark applications
- Submitting jobs to a cluster
- Monitoring and debugging Spark applications

14. PySpark on Cloud Platforms

- Running PySpark on AWS EMR
- Running PySpark on Google Cloud Dataproc
- Running PySpark on Azure HDInsight