PySpark

Duration: 20hrs

Module 1: Introduction to Big Data and Apache Spark

- **Overview of Big Data**
  - Understanding Big Data and its challenges
  - Big Data technologies landscape
- **Introduction to Apache Spark**
  - Spark vs. Hadoop: Understanding the differences.
  - Spark's ecosystem and components (Spark SQL, Spark Streaming, MLlib, GraphX)
- **Getting Started with PySpark**
  - Installing and setting up PySpark
  - Basic PySpark operations: Initializing SparkContext and SparkSession

Module 2: Core Concepts of PySpark

- **Resilient Distributed Datasets (RDDs)**
  - Fundamentals of RDDs
  - Creating RDDs and understanding partitioning
  - RDD operations: Transformations and Actions
- **DataFrames and Datasets**
  - Introduction to DataFrames and Datasets
  - Operations on DataFrames and Datasets
  - Interoperability between RDDs and DataFrames

Module 3: Data Processing with PySpark

- **Basic Data Manipulation**
  - Reading and writing data from various sources
  - Data selection, filtering, and aggregation
  - Handling missing data and data types
- **Advanced Data Processing**
  - Joins and complex operations
  - Window functions and group operations
  - UDFs (User Defined Functions) and their performance impact

Module 4: Structured Streaming

- **Introduction to Spark Streaming**

    - Understanding streaming data and Spark's streaming capabilities
    - Developing streaming applications
- **Structured Streaming Concepts**
    - Working with streaming DataFrames and Datasets
    - Stateful and stateless operations

Module 5: Machine Learning with PySpark

- **Introduction to MLlib**
    - Machine learning concepts
    - MLlib's data types and basic statistics
- **Building Machine Learning Models**
    - Supervised vs. unsupervised learning
    - Model evaluation and tuning
- **Advanced Machine Learning**
    - Pipelines and feature engineering
    - Scaling and deploying ML models

Module 6: Performance Tuning and Best Practices

- **Performance Optimization**
    - Understanding Spark's execution model
    - Tuning Spark jobs for performance
- **Best Practices in PySpark**
    - Writing efficient Spark code
    - Debugging and monitoring Spark applications

Module 7: Real-World PySpark Applications

- **Case Studies**
    - Analyzing real-world datasets
    - Building end-to-end Spark applications
- **Project Work**
    - Students will apply the concepts learned to a real-world data problem

Module 8: Conclusion

- **Review of Key Concepts**

Each module would be complete with lectures, hands-on labs, quizzes, and assignments to reinforce the concepts taught. The progression from basic to advanced topics ensures that learners build a solid foundation before tackling more complex problems, making this structure ideal for both beginners and those looking to deepen their understanding of PySpark.