Getting started with Apache Spark, Scala and HDInsight

***Duration** : 5 Days*

Prerequisites :

- ✓ Exposure to any Programming Language

Content Coverage :

Module 1: Introduction to Big Data and Analytics

- **What is Big Data?**
  - Characteristics (Volume, Velocity, Variety, Veracity)
  - Role of Big Data in today's business landscape
- **Challenges in Big Data Processing**
  - Storage, Scalability, Speed, and Security
- **Big Data Ecosystem Overview**
  - Tools and Technologies (Hadoop, Spark, Hive, Cassandra, Kafka)

Module 2: Introduction to Scala Programming for Big Data

- **Getting Started with Scala**
  - Installation and Setup of Scala
  - Introduction to Scala REPL and IDEs (IntelliJ IDEA, VSCode)
- **Scala Fundamentals**
  - Variables, Data Types, Control Structures (If/Else, Loops)
  - Functions and Methods
- **Object-Oriented Programming in Scala**
  - Classes, Objects, Traits, and Inheritance
- **Functional Programming in Scala**

    - Higher-order functions, Anonymous Functions, Immutability
    - Pattern Matching
- **Working with Collections in Scala**
    - Arrays, Lists, Sets, Maps, and Tuples
    - Operations on Collections: Filtering, Mapping, Reducing
- **Error Handling in Scala**
    - Exception Handling, Try, Option, Either

Module 3: Introduction to Apache Spark

- **What is Apache Spark?**
    - Spark vs. Hadoop: Why Spark for Big Data?
    - Spark Ecosystem: Spark Core, Spark SQL, Spark Streaming, MLlib, GraphX
- **Understanding Spark Architecture**
    - Driver, Executors, Cluster Managers (Standalone, YARN, Mesos)
- **Working with RDDs (Resilient Distributed Datasets)**
    - RDD Operations: Transformations (Map, Filter, FlatMap) and Actions (Collect, Reduce)
    - Fault tolerance in Spark (Lineage and Persistence)

Module 4: DataFrames and Datasets in Apache Spark

- **Introduction to DataFrames and Datasets**
    - Differences between RDDs, DataFrames, and Datasets
- **Creating DataFrames**
    - From structured data sources (CSV, JSON, Parquet)
    - Schema Inference and Explicit Schemas
- **Performing Operations on DataFrames**
    - Transformations (Select, Filter, GroupBy, Join)
    - Spark SQL for Structured Data Queries
    - Aggregations, Sorting, and Window Functions

Module 5: Advanced Spark Concepts

- **Spark Streaming for Real-Time Data Processing**
    - DStreams, Windowing Operations, Stateful Processing
    - Integration with Kafka for real-time data ingestion
- **Machine Learning with Spark MLlib**
    - Pipelines, Feature Engineering, Model Training (Regression, Classification, Clustering)
    - Building and Evaluating ML Models
- **Graph Processing with Spark GraphX**
    - Creating Graphs, Transformations, Graph Algorithms (PageRank, Connected Components)
- **Spark Performance Tuning**
    - Caching, Data Partitioning, Resource Management
    - Understanding Shuffling, Serialization, and Garbage Collection

Module 6: Introduction to Azure HDInsight

- **Overview of Azure HDInsight**
    - What is HDInsight? Use cases and architecture
    - Supported Services (Hadoop, Spark, Hive, Kafka, HBase, Storm, Cassandra)
- **Creating and Managing HDInsight Clusters**
    - Setting up a Spark Cluster on HDInsight
    - Monitoring and Scaling Clusters
    - Cost Optimization for HDInsight Clusters
- **HDInsight Security Features**
    - Secure Clusters, Role-Based Access Control (RBAC), and Encryption
    - Identity Management and Integration with Active Directory

Module 7: Spark on HDInsight

- **Running Apache Spark on HDInsight**
    - Submitting Spark jobs to HDInsight
    - Managing and Monitoring Spark Jobs
    - Working with DataFrames and RDDs in HDInsight

- **Integration with Azure Data Lake**
    - Connecting HDInsight to Azure Data Lake Storage (ADLS)
    - Loading and processing data from ADLS using Spark
- **Integrating HDInsight with Other Azure Services**
    - Using Azure Blob Storage, Data Factory, and Event Hubs with HDInsight
    - Connecting to SQL Databases and Data Warehouses

Module 8: Real-Time Analytics with Spark, HDInsight, and Cassandra

- **Building a Real-Time Analytics Pipeline**
    - Streaming Data Ingestion with Kafka or Event Hubs
    - Processing and Analyzing Streaming Data using Spark Streaming
    - Storing and Querying Results in Cassandra
- **Optimizing Real-Time Data Pipelines**
    - Ensuring Fault Tolerance and Data Integrity
    - Latency Optimization in Spark Streaming
    - Monitoring and Scaling Spark Streaming Jobs on HDInsight

Module 9: Security and Compliance in Big Data Systems

- **Securing Big Data on HDInsight**
    - Encryption at Rest and in Transit
    - Managing Access Control with Azure AD and HDInsight
    - Secure Connections to Data Sources and Services
- **Compliance with GDPR, HIPAA, and Data Privacy Laws**
    - Data Anonymization and Pseudonymization
    - Managing Consent and Handling Data Requests
    - Auditing and Logging Data Access
- **Disaster Recovery and Backups**
    - Backup Strategies for HDInsight Clusters
    - High Availability Configurations for Spark Jobs

Module 10 : Project : An Apache Spark End to End Data Engineering Project