Pyspark Programming with Databricks (5 days)

By Dr. Vishwanath Rao

Who Should Attend

PySpark Development training is suitable for IT professionals, software developers and data scientists who need to explore and analyze data more effectively
It is especially well-suited for people with no experience in Python programming – getting familiar with the language syntax and features can be difficult, but PySpark makes this much easier

Learning Objectives of PySpark Development

PySpark Development Training is designed to help students gain the necessary knowledge and skills to develop production-ready applications with Apache Spark.

Students will learn to develop applications with Scala and Python, understand how to scale and distribute data-intensive applications with Spark, and become familiar with the Spark core APIs.

The course will also focus on advanced topics such as machine learning and streaming, and provide insight into big data architectures and its role in enterprise data processing.

Students will gain proficiency in deploying and developing applications with Apache Spark and be able to confidently make data-driven decisions.

Course Contents

Section 1: Big Data Analytics introduction

- **Big Data overview**
- **Characteristics of Apache Spark**
- **Users and Use Cases of Apache Spark**
- **Job Execution Flow and Spark Execution**
- **Complete Picture of Apache Spark**
- **Why Spark with Python**
- **Apache spark Architecture**

- **Big Data Analytics in industry**

Section 2: Using Hadoop's Core: HDFS and MapReduce

- **HDFS: What it is, and how it works**
- **MapReduce: What it is, and how it works**
- **How MapReduce distributes processing**
- **HDFS commands**

Section 3: SparkDatabox Cloud Lab

- **How to access SparkDatabox cloud lab?**
- **Step by Step instruction to access cloud Big data Lab.**

Section 4: Data analytics lifecycle

- **Data Discovery**
- **Data Preparation**
- **Data Model Planning**
- **Data Model Building**
- **Data Insights**

Section 5: python 3.0 (Crash Course)

- **Environment Setup**
- **Decision Making**
- **Loops and Number**
- **Strings**
- **Lists**
- **Tuples**
- **Dictionary**
- **Date and Time**
- **Regex**
- **Functions**
- **Modules**

- **Files I/O**
- **Exceptions**
- **MultiThreading**
- **Set**
- **Lamda Function**

Section 6: PySpark

- **Introduction to SparkContext**
- **Environment Setup**
- **Spark RDD**
- **spark Caching**
- **Common Transformations and Actions**
- **Spark Functions**
- **Key-Value Pairs**
- **Aggregate Functions**
- **Working with Aggregate Functions**
- **Joins in Spark**
- **Spark DataFrame**

Section 7: Advanced Spark Programming

- **Spark Shared Variables**
- **Custom Accumulator**
- **Spark and Fault Tolerance**
- **Broadcast variables**
- **Numeric RDD Operations**
- **Per-Partition Operations**

Section 8: Running Spark jobs on Cluster

- **Spark Runtime Architecture**
- **Spark Driver**
- **Executors**

- Cluster Managers
- Connecting Spark To Different File System and Perform ETL ,(Extration Transformation and Loading)
- Connecting Spark To DataBases and Perform ETL (Extration Transformation and Loading)
- Spark StorageLevel
- Spark Serializers
- Spark-Submit and Cluster Explanation
- Performance Tuning

Section 9: PySpark Streaming at Scale

- Introduction to Spark Streaming
- PySpark Streaming with Apache Kafka
- Real-world Practical use cases
- Operations On Streaming Dataframes and Datasets
- Window Operations

Section 10: Real-world project training

- PySpark project environment setup
- Real-world PySpark project
- Project demonstration
- Expert evaluation and feedback

Section 11: You made it!!

- Spark Databox PySpark certification
- Interview preparation
- Mock interviews
- Resume preparation
- Knowledge sharing with industry experts
- Counseling to guide you to a right path in PySpark development career