PYSPARK COURSE CONTENTS

By Dr. Vishwanath Rao

PySpark Course Content

Introduction to the Basics of Python

- Explaining Python and Highlighting Its Importance
- Setting up Python Environment and Discussing Flow Control
- Running Python Scripts and Exploring Python Editors and IDEs

Sequence and File Operations

- Defining Reserve Keywords and Command Line Arguments
- Describing Flow Control and Sequencing
- Indexing and Slicing
- Learning the xrange() Function
- Working Around Dictionaries and Sets
- Working with Files

Functions, Sorting, Errors and Exception, Regular Expressions, and Packages

- Explaining Functions and Various Forms of Function Arguments
- Learning Variable Scope, Function Parameters, and Lambda Functions
- Sorting Using Python
- Exception Handling
- Package Installation
- Regular Expressions

Python: An OOP Implementation

- Using Class, Objects, and Attributes
- Developing Applications Based on OOP
- Learning About Classes, Objects and How They Function Together
- Explaining OOPs Concepts Including Inheritance, Encapsulation, and Polymorphism, Among Others

Debugging and Databases

- Debugging Python Scripts Using pdb and IDE
- Classifying Errors and Developing Test Units
- Implementing Databases Using SQLite
- Performing CRUD Operations

Introduction to Big Data and Apache Spark

- What is Big Data?
- 5 V’s of Big Data
- Problems related to Big Data: Use Case
- What tools available for handling Big Data?
- What is Hadoop?
- Why do we need Hadoop?
- Key Characteristics of Hadoop
- Important Hadoop ecosystem concepts
- MapReduce and HDFS
- Introduction to Apache Spark
- What is Apache Spark?
- Why do we need Apache Spark?
- Who uses Spark in the industry?
- Apache Spark architecture
- Spark Vs. Hadoop
- Various Big data applications using Apache Spark

Python for Spark

- Introduction to PySpark
- Who uses PySpark?
- Why Python for Spark?
- Values, Types, Variables
- Operands and Expressions
- Conditional Statements
- Loops
- Numbers
- Python files I/O Functions
- Strings and associated operations
- Sets and associated operations
- Lists and associated operations
- Tuples and associated operations
- Dictionaries and associated operations

Python for Spark: Functional and Object-Oriented Model

- Functions
- Lambda Functions
- Global Variables, its Scope, and Returning Values
- Standard Libraries
- Object-Oriented Concepts
- Modules Used in Python
- The Import Statements
- Module Search Path
- Package Installation Ways

Apache Spark Framework and RDDs

- Spark Components & its Architecture
- Spark Deployment Modes
- Spark Web UI
- Introduction to PySpark Shell
- Submitting PySpark Job
- Writing your first PySpark Job Using Jupyter Notebook
- What is Spark RDDs?
- Stopgaps in existing computing methodologies
- How RDD solve the problem?
- What are the ways to create RDD in PySpark?
- RDD persistence and caching
- General operations: Transformation, Actions, and Functions
- Concept of Key-Value pair in RDDs
- Other pair, two pair RDDs
- RDD Lineage
- RDD Persistence
- WordCount Program Using RDD Concepts
- RDD Partitioning & How it Helps Achieve Parallelization
- Passing Functions to Spark

PySpark SQL and Data Frames

- Need for Spark SQL
- What is Spark SQL
- Spark SQL Architecture
- SQL Context in Spark SQL
- User-Defined Functions
- Data Frames

- Interoperating with RDDs
- Loading Data through Different Sources
- Performance Tuning
- Spark-Hive Integration

Introduction to Apache Kafka and Flume

PySpark Streaming

- Introduction to Spark Streaming
- Features of Spark Streaming
- Spark Streaming Workflow
- StreamingContext Initializing
- Discretized Streams (DStreams)
- Input DStreams, Receivers
- Transformations on DStreams
- DStreams Output Operations
- Describe Windowed Operators and Why it is Useful
- Stateful Operators
- Vital Windowed Operators
- Twitter Sentiment Analysis
- Streaming using Netcat server
- WordCount program using Kafka-Spark Streaming

Introduction to PySpark Machine Learning (Self Paced)

- Introduction to Machine Learning- What, Why and Where?
- Use Case
- Types of Machine Learning Techniques
- Why use Machine Learning for Spark?
- Applications of Machine Learning (general)
- Applications of Machine Learning with Spark
- Introduction to MLlib
- Features of MLlib and MLlib Tools
- Various ML algorithms supported by MLlib
- Supervised Learning Algorithms
- Unsupervised Learning Algorithms
- ML workflow utilities