

Python and PySpark

**Duration- 7 days**

# Lab Pre-requisites

1. **Need Cloud VMs on cloud with 4 Core CPU for each participant**
2. **These VMs should have PySpark 3 installed**

# Notes:

1. **Each concept in the training will be followed by the exercises to understand the concepts by doing**
2. **There will be pre-assessment of the participants before the training starts.**

# INTRODUCTION TO THE BASICS OF PYTHON

* Explaining Python and highlighting its importance
* Setting up the Python environment and discussing flow control
* Running Python scripts and exploring Python editors and IDEs

# SEQUENCE & FILE OPERATIONS

* Defining reserve keywords and command-line arguments
* Describing flow control and sequencing
* Indexing and slicing
* Learning the range() function
* Working around dictionaries and sets
* Working with files

# FUNCTIONS, SORTING, ERRORS & EXCEPTIONS, REGULAR EXPRESSIONS, & 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 and objects and how they function together
* Explaining OOPs concepts, including inheritance, encapsulation, and polymorphism, among others

# DEBUGGING & 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 & APACHE SPARK

* What is Big Data?
* 5 Vs of Big Data
* Problems related to Big Data: Use Cases
* What are the tools available for handling Big Data?
* What is Hadoop?
* Why Hadoop?
* Key characteristics of Hadoop
* Important Hadoop ecosystem concepts
* MapReduce and HDFS
* Introduction to Apache Spark
* What is Apache Spark?
* Why 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, and 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
* Hands-on Exercise: Demonstrating loops and conditional statements, tuples: related operations, properties, etc., lists: operations, related properties, etc., sets: properties, associated operations, etc., and dictionaries: operations, related properties, etc.

# PYTHON FOR SPARK: FUNCTIONAL & OBJECT-ORIENTED MODEL

* Functions
* Lambda functions
* Global variables, their scope, and returning values
* Standard libraries
* Modules used in Python



* The import statement
* Module search path
* Package installation ways
* Hands-on Exercise: Lambda: features, options, syntax, and comparison with the functions, functions: syntax, return values, arguments, and keyword arguments, errors and exceptions: issue types and remediation, packages and modules: import options, modules, and sys path

# APACHE SPARK FRAMEWORK & RDDS

* Spark components and its architecture o Spark deployment modes
* Spark Web UI
* Introduction to PySpark Shell
* Submitting a PySpark job
* Writing your first PySpark job using Jupyter Notebook
* What are Spark RDDs?
* Stopgaps in existing computing methodologies
* How does an RDD solve the problem?
* What are the ways to create RDDs in PySpark?
* RDD persistence and caching
* General operations: Transformations, actions, and functions
* Concept of Key-Value pair in RDDs
* Other pair and two pair RDDs
* RDD lineage
* WordCount program using RDD concepts
* RDD partitioning and how it helps achieve parallelization
* Passing functions to Spark
* Hands-on Exercise: Building and running a Spark application, Spark Application Web UI, loading data into RDDs, saving the data through RDDs, RDD transformations, RDD actions and functions, RDD partitions, and the WordCount program using RDDs in Python

# PYSPARK SQL & DATAFRAMES

* Need for Spark SQL
* What is Spark SQL?
* Spark SQL architecture
* SQL Context in Spark SQL
* User-defined functions
* DataFrames
* Interoperating with RDDs
* Loading data through different sources
* Performance tuning
* Spark–Hive integration
* Hands-on Exercise: Spark SQL: Creating DataFrames, loading and transforming data through different sources, and Spark–Hive integration

# Slowly Changing Dimension Type 2 in Spark

* What’s an SCD?
* Implementation in Apache Spark
  + Pre-requisites
  + New Row
  + Deleted rows
  + Updated or rows which didn’t change

# Case Study

* Case study on Finance industry will be provided. Participants need to work on this case study offline and ask for doubts if any during the session. At the end of the training these case study will be submitted to the trainer for evaluation