**Databricks Using Spark (PySpark)**

**Apache Spark Training Overview**

Spark is a unique framework for big data analytics which gives one unique integrated API by developers for the purpose of data scientists and analysts to perform separate tasks. It supports a wide range of popular languages like **Python**, R, SQL, Java and Scala. Apache Spark main aim is to provide hands-on experience to create real-time **Data Stream Analysis** and large-scale learning solutions for data scientists, data analysts and software developers.

**Apache Spark Training Objectives**

* Apache Spark Architecture How to use Spark with Python (PySpark) to build data pipelines.

**Pre-requisites of the Course**

* Basic knowledge of object-oriented programming is enough Knowledge of Python will be an added advantage
* Learners who have basic knowledge on Database, [SQL](http://52.7.23.29/sql-server-training/) Query will be an added advantage for Learning this Course

**Who should do the course**

* Developers, Architects, IT Professionals
* Aspiring Software Engineers, Data scientists, and Analysts

**Python Basics (1–2 weeks)**

1. **Introduction to Python**
   * Python syntax and semantics
   * Variables, data types (strings, integers, floats, Booleans)
   * Conditional statements (if, elif, else)
   * Loops (for, while)
2. **Functions**
   * Defining functions
   * Arguments and return values
   * Lambda functions
   * Recursion
3. **Data Structures**
   * Lists, tuples, sets, and dictionaries
   * List comprehensions
   * Iterators and Generators
4. **Error Handling and Exceptions**
   * Try/except blocks
   * Raising exceptions
   * Debugging techniques
5. **Object-Oriented Programming**
   * Classes and Objects
   * Instance and class variables
   * Methods and magic methods (\_\_init\_\_, \_\_str\_\_, etc.)

**Introduction to Big Data & PySpark (2 weeks)**

1. **What is Big Data?**
   * Introduction to Big Data technologies
   * Hadoop ecosystem vs Spark ecosystem
   * Distributed computing fundamentals
2. **Apache Spark Basics**
   * What is Apache Spark?
   * Spark components (Core, SQL, MLlib, GraphX)
   * Setting up Spark locally and in Databricks
3. **Spark RDD (Resilient Distributed Dataset)**
   * Understanding RDDs
   * Transformations (map, filter, flatMap)
   * Actions (collect, count, save)
   * Persisting RDDs
4. **PySpark DataFrames**
   * Introduction to DataFrames in PySpark
   * Creating DataFrames
   * DataFrame transformations and actions
   * Operations on DataFrame (select, filter, groupBy, join)
5. **Spark SQL**
   * SQL queries in Spark
   * Registering DataFrames as SQL tables
   * Using spark.sql() to run SQL queries

**Working with Databricks (1 week)**

1. **Introduction to Databricks**
   * Setting up a Databricks workspace
   * Databricks notebooks
   * Cluster setup and management
   * Running jobs on Databricks
2. **Databricks Notebooks for Data Analysis**
   * Basic notebook operations (cells, markdown, etc.)
   * Visualizations in Databricks
   * Using Databricks notebooks for collaboration
3. **Integrating PySpark with Databricks**
   * Writing PySpark code in Databricks notebooks
   * Running PySpark jobs on Databricks clusters
   * Accessing data from Databricks file system
4. **Using Databricks with Cloud Storage**
   * Working with AWS S3, Azure Blob Storage, and GCP Storage in Databricks
   * Data loading and saving (CSV, Parquet, JSON)

**Advanced PySpark and Databricks (1 week)**

1. **Advanced PySpark Topics**
   * Window functions
   * Advanced aggregations
   * Handling skewed data
   * PySpark performance tuning
2. **Streaming with PySpark**
   * Introduction to Spark Streaming
   * Real-time data processing using PySpark Streaming
   * Kafka integration with PySpark
3. **Optimizing Spark Jobs**
   * Caching and persisting RDDs/DataFrames
   * Partitioning and shuffling
   * Tuning Spark jobs for performance
4. **Advanced Databricks Features**
   * Databricks jobs and workflows
   * Scheduling notebooks and jobs
   * Collaborating with team members using Databricks

Name: Bade Vamshikrishna

Mail: [vamshilrishnabade1003@gmail.com](mailto:vamshilrishnabade1003@gmail.com)

LinkedIn: <https://www.linkedin.com/in/vamshikrishnabade/>