CREDO SYSTEMZ
Simplifying IT

Pyspark Training Course Content

CHAPTER 1: Big Data

- Big Data Overview
- Processing Frameworks
- Programming Languages & Databases
- Lambda Architecture

CHAPTER 2: Python

- Python Introduction
- Python Object Tour
- Functional programming
- Analytics using Dataframes

CHAPTER 3: Lowel level APIs

- RDDs
- Creating & Manipulating RDDs
- Transformations & Actions
- Connecting to Data Sources

CHAPTER 4: Advanced RDDs

- Key-value RDDs
- Aggregations
- Joins
- Custom partitioning
- Distributed shared variables

CHAPTER 5: Structured APIs

- DataFrames
- Spark data types

- Operations : Filters, Column manipulations, Unions, Joins, Aggregations, Sorting
- I/O operations with different types of data sources (TEXT,CSV,JSON,PARQUET,DB)
- Spark SQL
- Datasets

CHAPTER 6: Productionizing applications

- How Spark runs on a cluster
- Submitting Spark applications
- Deployment Modes
- Monitoring & Debugging
- Performance Tuning

CHAPTER 7: Structured Streaming

- Fundamentals on streams
- Structured streaming
- Event time and stateful processing
- Real time streaming application

CHAPTER 8: Machine learning using Spark

- Introduction
- Preprocessing and Feature Engineering
- Classification
- Regression
- Recommendation
- Unsupervised learning

CHAPTER 9: Real-time Projects