Apache Spark Training

Apache Spark is a big data processing framework and its popularity lies in the fact that it is fast, easy to use and offers sophisticated solutions to data analysis. Its built-in modules for streaming, machine learning, SQL, and graph processing make it useful in diverse Industries like Banking, Insurance, Retail, Healthcare, and Manufacturing.

Course Objective:

- Understand the architecture of Spark and explain its business use cases
- Distribute, store, and process data using RDDs in a Hadoop cluster
- Use Spark SQL for querying DBs
- Write, configure, and deploy Spark applications on a cluster
- Use the Spark shell for interactive data analysis
- Process and query structured data using Spark SQL

Course Audience:

Professionals aspiring to learn the basics of Big Data Analytics using Spark Framework and become a Spark Developer. In addition, it would be useful for Analytics Professionals and ETL developers as well.

Course Prerequisites:

Those wishing to take the Apache Spark certification training should have a fundamental knowledge of any programming language and a basic understanding of any database, SQL, and query language for databases. Working knowledge of Linux- or Unix-based systems is also beneficial.

Introduction to Spark:

- What is Spark?
- Spark Overview
- Setting up environment
- Build a simple Spark project with Eclipse & Maven
- Using Spark Shell

Spark Basics:

- Resilient Distributed Datasets (RDDs)
- Spark Context
- Spark Ecosystem
- In-Memory Computations in Spark

Working with RDDs:

- Creating, Loading and Saving RDD
- Transformations on RDD
- Actions on RDD
- Key-Value Pair Transformation on RDDs
- RDD Partitioning
- RDD Persistence

Writing and Deploying on Cluster:

- Spark Applications vs. Spark Shell
- Spark Runtime Architecture
- Creating Spark Context
- Building a Spark Application
- Deploying Spark Applications using Spark-Submit

Spark Job Execution :

- RDD Lineage
- Jobs, Stages and Tasks
- Partition and Shuffles
- Data Locality
- Join with or without Partitioner, stages and tasks, etc
- Spark Web UI

Spark SQL:

- Overview on Hive
- Spark SQL Architecture
- SparkSession in Spark SQL
- Working with DataFrames
- Integrating Spark SQL with Hive
- Integrating Spark SQL with JDBC Sources (MySQL)
- Integrating Spark SQL with NoSQL DB (Cassandra)
- Handling CSV, JSON and Parquet File Formats
- Loading and Saving Data

Spark Streaming:

- Spark Streaming Architecture
- Spark Streaming Transformations
- Rolling Window and Check pointing
- Integrating Spark with Kafka Streaming Data
- Structured Streaming
- Integrating Spark with Twitter Streaming Data
- Spark Streaming Performance Considerations

Spark MLlib:

- What is Machine Learning?
- ML library for Spark
- ML Concepts and Algorithms
- Typical Steps in ML Pipeline – Executors and Transformers
- ML using Pipelines and DataFrames
- Recommendation Engine – Practical Use Case

Spark GraphX:

- Overview of GraphX
- Components of GraphX – VertexRDD, EdgeRDD and Triplets
- Develop simple application with GraphX
- Transformations on GraphX
- Hands on – PageRank, TriangleCount Algorithms
- Common Spark Use-cases

Performance Tuning and Debugging:

- Shared Variables: Broadcast Variables
- Shared Variables: Accumulators
- Common Performance Issues
- Performance Tuning Tips
- Spark WebUI
- Monitoring Driver and Executor Logs