Hadoop Developer with Spark

Course Outcomes:

- Distribute, store, and process data in a Hadoop cluster
- 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 and Hive Query Language
- Understand a wide variety of learning algorithms and Build an end-to-end Machine Learning Model with MLlib in pySpark.
- Use Spark Streaming to process a live data stream

What to Expect

This course is designed for developers and engineers who have programming experience, but prior knowledge of Hadoop and/or Spark is not required.

- Apache Spark examples and hands-on exercises are presented in Scala and Python. The ability to program in one of those languages is required.
- Basic familiarity with the Linux command line is assumed
- Basic knowledge of SQL is helpful

Course Duration: 40 Hours

Module 1

Introduction to Apache Hadoop

and the Hadoop Ecosystem

• Apache Hadoop Overview
• Data Ingestion and Storage
• Data Processing
• Data Analysis and Exploration
• Other Ecosystem Tools
• Introduction to the Hands-On Exercises

Module 2

Apache Hadoop File Storage

• Apache Hadoop Cluster Components
• HDFS Architecture
• Using HDFS

Module 3

Distributed Processing on an Apache Hadoop Cluster

• YARN Architecture

• Working With YARN

Module 4
Apache Spark Basics
- What is Apache Spark?
- Starting the Spark Shell
- Using the Spark Shell
- Getting Started with Datasets and DataFrames
- DataFrame Operations

Module 5
Working with DataFrames and Schemas
- Introduction to DataFrames
- Exercise: Introducing DataFrames
- Exercise: Reading and Writing DataFrames
- Exercise: Working with Columns
- Exercise: Working with Complex Types
- Exercise: Combining and Splitting DataFrames
- Exercise: Summarizing and Grouping DF
- Exercise: Working with UDFs
- Exercise: Working with Windows
- Eager and Lazy Execution

Module 6
Analyzing Data with DataFrame Queries
- Querying DataFrames Using Column Exp.
- Grouping and Aggregation Queries
- Joining DataFrames

Module 7
Introduction to Apache Hive
- About Hive
- Transforming data with Hive QL

Module 8
Working with Apache Hive
- Exercise: Working with Partitions
- Exercise: Working with Buckets
- Exercise: Working with Skew
- Exercise: Using Serdes to Ingest Text Data
- Exercise: Using Complex Types to Denormalize Data

Module 9
Hive and Spark Integration
- Hive and Spark Integration

• Exercise: Spark Integration with Hive

Module 10
RDD Overview
• RDD Overview
• RDD Data Sources
• Creating and Saving RDDs
• RDD Operations

Module 11
Transforming Data with RDDs
• Writing and Passing Transformation Functions
• Transformation Execution
• Converting Between RDDs and DataFrames

Module 12
Aggregating Data with Pair RDDs
• Key-Value Pair RDDs
• Map-Reduce
• Other Pair RDD Operations

Module 13
Querying Tables and Views with Apache Spark SQL
• Querying Tables in Spark Using SQL
• Querying Files and Views
• The Catalog API
• Comparing Spark SQL, Apache Impala,
and Apache Hive-on-Spark

Module 14
Working with Datasets in Scala
• Datasets and DataFrames
• Creating Datasets
• Loading and Saving Datasets
• Dataset Operations

Module 15
Writing, Configuring, and Running Apache Spark Applications
• Writing a Spark Application
• Building and Running an Application
• Application Deployment Mode
• The Spark Application Web UI
• Configuring Application Properties

Module 16
Distributed Processing
- Review: Apache Spark on a Cluster
- RDD Partitions
- Example: Partitioning in Queries
- Stages and Tasks
- Job Execution Planning
- Example: Catalyst Execution Plan
- Example: RDD Execution Plan

Module 17
Distributed Processing Challenges
- Shuffle
- Skew
- Order

Module 18
Distributed Data Persistence
- DataFrame and Dataset Persistence
- Persistence Storage Levels
- Viewing Persisted RDDs

Module 19
Machine Learning with Spark ML
- Common Apache Spark Use Cases
- Iterative Algorithms in Apache Spark : Machine Learning, Graph Processing
- **Introduction to MLlib- Various ML algorithms supported by Mlib**
- **ML model with Spark ML**
- Exercise: **Implement Linear regression**
- Exercise: **Implement logistic regression**
- Exercise: **Implement Random Forest**
- Exercise: **Implement** k-means

Module 20
Apache Spark Streaming: Introduction to DStreams
- Apache Spark Streaming Overview
- Example: Streaming Request Count
- DStreams
- Developing Streaming Applications

Module 21
Apache Spark Streaming: Processing Multiple Batches
- Multi-Batch Operations
- Time Slicing

- State Operations
- Sliding Window Operations
- Preview: Structured Streaming

Module 22

Apache Spark Streaming: Data Sources

- Streaming Data Source Overview
- Apache Flume and Apache Kafka Data Sources
- Example: Using a Kafka Direct Data Source