BigData - Hadoop / Spark – Course Outline

1 Duration

- 3 Days

2 Objectives

At end of this workshop, participants will able to :

- Get an overall understanding of key technologies involved in Big Data space and Spark Ecosystem
- Understand fundamentals of Hadoop and Spark, Spark vs MR and YARN
- Get knowledge on Hive, Sqoop and Flume
- Get knowledge on Spark Core, Spark SQL
- Get a feel of some of the technologies in action with hands-on and real time use cases
- Troubleshoot and fine tune Hadoop/Spark components and learn usage patterns and best practices

3 Audience

Big Data Developers, Enterprise Warehouse Professionals and QA Professionals who wanted to get themselves familiarized with BigData technologies.

4 Pre-requisite

- Familiarity with Distributed Computing
- Programming knowledge on Java/Scala/Python
- Good knowledge on Unix commands
- Good knowledge on SQL

5 Hardware & Network Requirements

- Desktop with minimum 8GB RAM (16 GB Recommended)
- Good Internet connection (minimum 1 mbps for each participant)

6 Software Requirements

- Windows / Linux OS
- Oracle VirtualBox 5.2 and above (to run Hadoop/Spark Image being shared)

7 Outline

7.1 Day 1

1) Big Data & Hadoop Overview

a) Big Data Overview
b) Hadoop Overview
c) Hadoop Key Characteristics
d) Hadoop Ecosystem
e) Hadoop Core Components
f) Hadoop Setup, Configuration and Data Loading
g) Hadoop 1.x vs 2.x vs 3.x
h) **Labs/Demo:** How to start/stop and configure Hadoop components

2) HDFS Fundamentals and Internals

a) HDFS Architecture
b) Components of HDFS - Name Node, Secondary Name Node, Data Node
c) HDFS File Write Anatomy
d) HDFS File Read Anatomy
e) Hadoop File Formats and Compression Techniques
f) **Labs/Demo:** Load weather dataset into HDFS and perform different operations using Hadoop CLI

3) Hadoop MapReduce

a) MapReduce Framework, Anatomy and Flow
b) MapReduce concepts - Splits, Mappers, Reducers, Partitioners, Combiners and Counters
c) Input / Output File Formats
d) Map side join / Reduce side join
e) Distributed cache
f) **Labs/Demo:** Write MR program to load, process and analyse sample weather dataset.

7.2 Day 2

1) Hive

a) Introduction to Hive
b) Overview of Hive2
c) Hive Setup, Configuration and Commands
d) Hive Components, Architecture, Metastore
e) Hive Data Types
f) Hive Data Models
g) Hive Managed Tables, External Tables, Partitioned Tables, Clustered Tables concepts
h) **Labs/Demo:** Write Hive scripts to load, process and analyse sample weather dataset

2) Sqoop and Flume

a) Introduction
b) Setup and Configuration
c) Examples to import / export data
d) **Labs/Demo:** Write Sqoop program to extract and load weather dataset from database.
Write Flume program to extract and load the weather dataset from file/streaming source

3) **Introduction to Spark**
   a) Spark Overview
   b) MR vs Spark
   c) Spark Installation and Modes of Operation
   d) Spark Fundamentals, Architecture, Components
   e) Spark on YARN
   f) Spark Context
   g) RDD Fundamentals
   h) Job server
   i) Spark Programming with Java/Scala/Python

7.3 Day 3

1) **Spark Core**
   a) Transformations and Actions in RDD
   b) RDD API in Detail
   c) RDD Persistence
   d) Types of RDD
   e) RDD Partitioning
   f) Accumulators and Broadcast Variables
   g) Creating RDD from Different File Formats
   h) **Labs/Demo:** Write Spark program to load, process and analyse sample weather dataset

2) **Introduction to Spark SQL**
   a) Spark SQL Overview
   b) Data Frames in Detail
   c) Creating Data Frames
   d) Transformations and Actions on Data Frames
   e) Various Spark SQL Operations
   f) Data Sources
   g) Spark Schedulers
   h) **Labs/Demo:** Write Spark SQL program to load, process and analyse sample weather dataset

3) **Hadoop/Spark Distributions and Latest Trends in Big Data Analytics**
   a) Overview of various Hadoop/Spark distributions
   b) Overview of Cluster Administration, Troubleshooting and Monitoring
   c) Overview of Latest Trends in Big Data Analytics space