Hadoop and PySpark – Course Outline

1 Duration

- 40 Hours

2 Objectives

At end of this workshop, participants will able to :

- Get an overall understanding of fundamentals and internals of Hadoop and Spark Ecosystem
- Get knowledge on HDFS, Map Reduce, Hive, Sqoop and HBase
- Get knowledge on Spark Core and PySpark programming
- Get understanding on Spark SQL
- Get basic understanding on Spark Streaming
- Get a feel of some of the technologies in action with hands-on and real time use cases

3 Audience

Developers, Enterprise Data Warehouse Professionals, QA Professionals or whoever wants to get themselves familiarized with Hadoop and Spark and technologies around it.

4 Pre-requisite

- Good Programming knowledge in Python
- Good knowledge on Unix commands
- Good knowledge on SQL

5 Hardware & Network Requirements

- Desktop/Laptop with minimum 8GB RAM (Recommended 16 GB)
- Open Internet connection (minimum 2 mbps per user)

6 Software Requirements

- Windows / Linux / Mac OS (with local admin access)
- Oracle VirtualBox (to practice in local machine)
- Cloud Labs to be provided during training for Labs

* Pre-configured image with all required softwares or setup instructions to be shared during training for labs.

7 Outline

Day 1

Module-1: Introduction to BigData and Hadoop

- What is Big Data?
- Challenges of Big Data World
- Big Data in Industry (Use cases)
- Limitations of traditional BI architecture
- What is Hadoop?
- Hadoop Key Characteristics
- Hadoop Architecture
- Hadoop Ecosystem
- Hadoop Core Components
- Hadoop Setup and Configuration
- Hadoop 1.x vs Hadoop 2.x vs Hadoop 3.x

Module-2: HDFS Internals

- HDFS Architecture
- Components of HDFS - Name Node, Secondary Name Node, Data Node
- HDFS File Write Anatomy
- HDFS File Read Anatomy
- Hadoop File Formats and Compression Techniques

Module-3: Map Reduce and YARN

- MapReduce Framework, Anatomy and Flow
- MapReduce concepts - Splits, Mappers, Reducers, Partitioners, Combiners and Counters
- YARN Architecture
- Resource Manager
- Node Manager
- Application Master
- Scheduler

Day 2

Module-4: Hive

- Introduction to Hive
- Overview of Hive2
- Hive Setup, Configuration and Commands
- Hive Components, Architecture, Metastore
- Hive Data Types
- Hive Data Models
- Hive Managed Tables, External Tables, Partitioned Tables, Clustered Tables concepts
- Hive SQL - Basics, DDL / DML Operations, Queries with Aggregation, Grouping, Sorting, Clustering, Union, Joins
- Writing Hive Queries to do data analysis
- Hive UDFs and UDAFs

Day 3

Module-5: HBase

- Introduction to HBase
- HBase Setup and Configuration
- HBase Components and Architecture
- Using the HBase Shell
- HBase General Commands
- HBase Schema Design
- HBase Data Model
- Create and Manage HBase tables
- Load data into HBase tables
- Query data from HBase tables

Module-6: Sqoop

- Introduction to Sqoop
- Overview of Sqoop2
- Sqoop Setup and Configuration
- Sqoop examples to import / export data

Module-6: Introduction to Spark

- Spark Overview
- MR vs Spark
- Spark Features
- Benefits and Limitations
- PySpark Overview
- PySpark programming

Day 4

Module-7: Spark Setup and Fundamentals

- Spark Installation and Modes of Operation
- Spark Fundamentals, Architecture, Components
- Spark on YARN
- Spark Context
- Spark Shell
- Job Server
- Spark 2.x vs Spark 3.x

Module-8: Spark Core

- RDD: The foundation of Spark
- Creating RDDs from different types of files
- Creating RDDs from another RDDs
- RDD operations, Actions and Transformations
- Different Types of RDDs
- Joins using RDD
- RDD Persistence and RDD Partitioning
- RDD Lineage and DAG
- Broadcast variables and Accumulators
- Connecting to Different Sources with Spark
- Spark programming with PySpark

Day 5

Module-9: Introduction to Spark SQL

- Spark SQL - Structured Data Processing
- Spark SQL integration with Hive
- SQL Context
- Data Frames in Detail
- Creating Data Frames
- Convert RDD to DataFrame using DataFrmae API for query data
- Transformations and Actions on Data Frames
- Various Spark SQL Operations
- Working with different Data Sources / File Formats
- Developing Spark SQL (Data Frames) Applications with PySpark
- Spark SQL integration with Hive

Module-10: Introduction to Spark Streaming

- Spark Streaming - Real time Data Processing
- Spark vs Storm
- Dstreams and Micro Batch
- Windowing Concept
- Dstreams Actions and Transformations
- Window Level Actions and Transformations
- Structured Streaming API
- Dstreams vs Structured Streaming
- Stream Processing with Structured Streaming using DataFrames
- Developing Spark Streaming (Dstreams) Applications with PySpark
- Developing Spark Streaming (Structured Streaming) Applications with PySpark