

**BDHS** 

# Big Data with Hadoop and Spark







## Big Data with Hadoop and Spark

BDHS - Version: 3



#### **Description:**

In the first training day, the course will introduce developers to the Hadoop ecosystem focusing on HDFS, Hive and HBase.

In the next two training days, the course will cover Spark including: Spark Streaming, Graphs and Machine Learning.

#### Intended audience:

Developers who wish to obtain the skills and knowledge to develop big data solutions with Hadoop and Spark.

### **Prerequisites:**

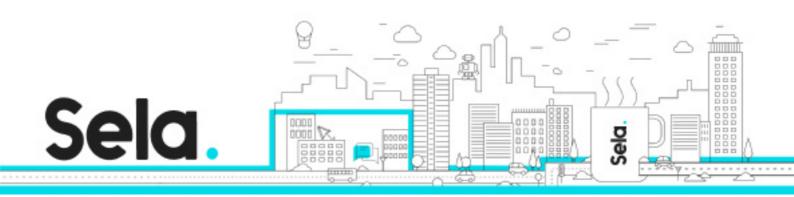
Basic knowledge of programming in Python

### **Objectives:**

Distributed programming with Spark
Using Machine learning algorithms
Storing and analyzing data using HBase and Hive

### **Topics:**

Hadoop (day 1)



- Hadoop Introduction
  - □ Big Data Analytics
  - <sup>o</sup> What is and Why Hadoop
  - Comparing Hadoop with Other Technologies
  - <sup>⁰</sup> Hadoop Architecture
  - <sup>⁰</sup> Hadoop Ecosystem
  - <sup>9</sup> Hadoop Usage Examples
- Hadoop HDFS
  - <sup>⁰</sup> What Is and Why HDFS
  - <sup>⁰</sup> HDFS Architecture
  - □ HDFS Features
  - <sup>⁰</sup> HDFS Commands
  - <sup>⁰</sup> HDFS Web UI
  - <sup>⁰</sup> Hue Web UI
- Hadoop HBase
  - What is and Why HBase?
  - HBase Shell

  - <sup>⁰</sup> HBase Data Model

  - <sup>o</sup> Accessing data in the HBase table using Pig and Hive
  - <sup>⁰</sup> Access HBase table via REST/Thrift
- Hadoop Hive
  - <sup>o</sup> What Is and Why Hive?
  - <sup>⁰</sup> Hive Architecture
  - <sup>⁰</sup> HiveQL
  - Physical Layout
  - <sup>⁰</sup> Loading Data into Hive Tables
  - Partitions
  - <sup>⁰</sup> Joining
  - <sup>⁰</sup> Buckets



### Spark Workshop (day 2 & 3)

- Why Spark?
- Basic Concepts
- Running on Clusters
- Spark SQL
- Spark Structured Streaming
- Spark ML
- Spark Performance Topics