## Course duration

4 days

## **Course Benefits**

- How the Apache Hadoop ecosystem fits in with the data processing lifecycle
- How data is distributed, stored, and processed in a Hadoop cluster
- · How to write, configure, and deploy Apache Spark applications on a Hadoop cluster
- How to use the Spark shell and Spark applications to explore, process, and analyze distributed data
- How to query data using Spark SQL, DataFrames, and Datasets
- How to use Spark Streaming to process a live data stream

## **Course Outline**

- 1. Introduction to Apache Hadoop and the Hadoop Ecosystem
  - 1. Apache Hadoop Overview
  - 2. Data Processing
  - 3. Introduction to the Hands-On Exercises
- 2. Apache Hadoop File Storage
  - 1. Apache Hadoop Cluster Components
  - 2. HDFS Architecture
  - 3. Using HDFS
- 3. Distributed Processing on an Apache Hadoop Cluster
  - 1. YARN Architecture
  - 2. Working With YARN
- 4. Apache Spark Basics
  - 1. What is Apache Spark?
  - 2. Starting the Spark Shell
  - 3. Using the Spark Shell
  - 4. Getting Started with Datasets and DataFrames
  - 5. DataFrame Operations
- 5. Working with DataFrames and Schemas
  - 1. Creating DataFrames from Data Sources
  - 2. Saving DataFrames to Data Sources
  - 3. DataFrame Schemas
  - 4. Eager and Lazy Execution
- 6. Analyzing Data with DataFrame Queries
  - 1. Querying DataFrames Using Column Expressions
  - 2. Grouping and Aggregation Queries
  - 3. Joining DataFrames

- 7. RDD Overview
  - 1. RDD Overview
  - 2. RDD Data Sources
  - 3. Creating and Saving RDDs
  - 4. RDD Operations
- 8. Transforming Data with RDDs
  - 1. Writing and Passing Transformation Functions
  - 2. Transformation Execution
  - 3. Converting Between RDDs and DataFrames
- 9. Aggregating Data with Pair RDDs
  - 1. Querying Tables in Spark Using SQL
  - 2. Querying Files and Views
  - 3. The Catalog API
  - 4. Comparing Spark SQL, Apache Impala, and Apache Hive-on-Spark
- 10. Querying Tables and Views with SQL
  - 1. Querying Tables in Spark Using SQL
  - 2. Querying Files and Views
  - 3. The Catalog API
- 11. Working with Datasets in Scala
  - 1. Datasets and DataFrames
  - 2. Creating Datasets
  - 3. Loading and Saving Datasets
  - 4. Dataset Operations
- 12. Writing, Configuring, and Running Spark Applications
  - 1. Writing a Spark Application
  - 2. Building and Running an Application
  - 3. Application Deployment Mode
  - 4. The Spark Application Web UI
  - 5. Configuring Application Properties
- 13. Spark Distributed Processing
  - 1. Review: Apache Spark on a Cluster
  - 2. RDD Partitions
  - 3. Example: Partitioning in Queries
  - 4. Stages and Tasks
  - 5. Job Execution Planning
  - 6. Example: Catalyst Execution Plan
  - 7. Example: RDD Execution Plan
- 14. Distributed Data Persistence
  - 1. DataFrame and Dataset Persistence
  - 2. Persistence Storage Levels
  - 3. Viewing Persisted RDDs
- 15. Common Patterns in Spark Data Processing
  - 1. Common Apache Spark Use Cases
  - 2. Iterative Algorithms in Apache Spark
  - 3. Machine Learning
  - 4. Example: k-means
- 16. Introduction to Structured Streaming

- 1. Apache Spark Streaming Overview
- 2. Creating Streaming DataFrames
- 3. Transforming DataFrames
- 4. Executing Streaming Queries
- 17. Structured Streaming with Apache Kafka
  - 1. Overview
  - 2. Receiving Kafka Messages
  - 3. Sending Kafka Messages
- 18. Aggregating and Joining Streaming DataFrames
  - 1. Streaming Aggregation
  - 2. Joining Streaming DataFrames
- 19. Conclusion
  - 1. Message Processing with Apache Kafka
  - 2. What Is Apache Kafka?
  - 3. Apache Kafka Overview
  - 4. Scaling Apache Kafka
  - 5. Apache Kafka Cluster Architecture
  - 6. Apache Kafka Command Line Tools

## **Class Materials**

Each student will receive a comprehensive set of materials, including course notes and all the class examples.

Class Prerequisites

Experience in the following *is required* for this Hadoop class:

- The ability to program in Scala or Python is required.
- · Basic familiarity with the Linux command line.

Experience in the following would be useful for this Hadoop class:

Basic knowledge of SQL.