RX-M - Apache Hadoop & Apache Spark for Data Science

Code: CN2-DS-HS
Length: 3 days
URL: View Online

Data science is an interdisciplinary field focused on developing insights from structured and unstructured data. Apache Hadoop and Apache Spark are some of the most important and most used tools in the data science field. This course will teach attendees how to use Apache Hadoop and Apache Spark to solve sophisticated data science problems, producing valuable insights in a wide range of scenarios.

Day one focuses on data science basics, including data acquisition, scrubbing and manipulation, as well as a general overview of data science applications and the analytics and machine learning processes typically used. A number of practical use cases are examined during class and lab sessions.

Day two focuses on Apache Hadoop and its ecosystem along with the types of data science applications typically handled by the Hadoop platform. The course outlines the statistical methods used to produce actionable business insights with Map Reduce, Python, Pig, Mahaut and other tools.

Day three begins with an overview of the Apache Spark platform and its machine learning library, MLlib. Attendees will learn how to perform entity ranking, implement recommendation engines and perform other common data science tasks using Spark batch, streaming, graph and machine learning capabilities. Upon course completion attendees will have a clear understanding of data science, its typical use cases and how data science is performed using a range of tools in the Apache open source ecosystem.

Skills Gained

- This course is designed to provide attendees with a comprehensive introduction to data science with Apache Spark and Apache Hadoop.

Who Can Benefit

- Application developers, analysts and data scientists

Prerequisites

- Each attendee will require the ability to run a 64 bit virtual machine (provided with the course). Basic Linux command line skills are valuable but not required.

Course Details

Apache Hadoop & Apache Spark for Data Science

- Day 1 - Data Science

1. Data Science Overview
2. Structured and Unstructured Data
3. Data Acquisition and Transformation
4. Data Analysis and Machine Learning

- Day 2 - Apache Hadoop

1. Map Reduce Fundamentals
2. Common Hadoop use cases
3. Machine Learning with Mahaut
4. NLTK and Natural Language Processing

- Day 3 - Apache Spark

1. Apache Spark Overview
2. Working with MLlib
3. Spark Streaming
4. Moving applications to production

Schedule (as of 4)

Date	Location