



Data Sheet

## Applying Data Science using Apache Hadoop

“Applying Data Science using Hadoop” covers Data Science principles and techniques through lecture and hands-on experience. During this two-day class, students will experience a hands-on learning environment to experience the processes and practice of data analysis with Hadoop and the R statistical language with the outcome of implementing a recommender solution with R and Mahout.

### Duration:

2 days

### Prerequisites:

Students must have basic computer skills, basic knowledge in statistics and a basic understanding of programming or scripting. Prior experience with Hadoop, Mahout, or R, although helpful, is not required.

### Target Audience:

Architects, software developers, analysts and data scientists who need to understand how to apply data science to large datasets with Hadoop.

### Format:

2 days of mixed lecture and hands-on labs

### Course Objectives

- Understand the basics of Data Science
- Understand the basics of machine learning
- Learn about Hadoop and its relation to Data Science
- Learn the basics of the R statistics language from Revolution Analytics
- Understand recommender systems
- Implement a recommender system with R statistics language
- Implement a recommender system with Hadoop (using Mahout)

## Agenda

### Day 1

- Overview
- Why Data Science?
- What is Data Science?
- Hadoop and Data Science
- The Process of Data Analysis
- Data and Functions in the R
- Data Analysis Using R

### Day 2

- Introduction to Machine Learning
- Recommender Systems
- Using a Sparse Matrix in R
- Recommender Algorithm with R
- Implement a Recommender System with Mahout
- Taking Data Science to Production
- Where to Learn More About Data Science

## Lab Content

- Hands on setup of solution environment
- Defining the problem
- Fundamentals of R
- Data analysis using R
- Creating the user/item matrix
- Using recommenderlab with R
- Running Mahout with Hadoop
- Mahout ALS & Evaluation
- Data product design diagram



**Hortonworks University** is your expert source for Apache Hadoop training and certification. Public and private on-site courses are available for developers, administrators, data analysts and other IT professionals involved in implementing big data solutions. Classes combine presentation material with industry-leading hands-on labs that fully prepare students for real-world Hadoop scenarios.



### About Hortonworks

Hortonworks is a leading commercial vendor of Apache Hadoop, the preeminent open source platform for storing, managing and analyzing big data. Our distribution, Hortonworks Data Platform powered by Apache Hadoop, provides an open and stable foundation for enterprises and a growing ecosystem to build and deploy big data solutions. Hortonworks is the trusted source for information on Hadoop, and together with the Apache community, Hortonworks is making Hadoop more robust and easier to install, manage and use. Hortonworks provides unmatched technical support, training and certification programs for enterprises, systems integrators and technology vendors.

**US:** 1.855.846.7866  
**International:** 1.408.916.4121  
**[www.hortonworks.com](http://www.hortonworks.com)**

3460 West Bayshore Road  
Palo Alto, CA 94303 USA