**Advanced Analytics with Spark: Patterns for Learning from Data at Scale (2015)**

In this practical book, four Cloudera data scientists present a set of self-contained patterns for performing large-scale data analysis with Spark. The authors bring Spark, statistical methods, and real-world data sets together to teach you how to approach analytics problems by example.

You’ll start with an introduction to Spark and its ecosystem, and then dive into patterns that apply common techniques—classification, collaborative filtering, and anomaly detection among others—to fields such as genomics, security, and finance. If you have an entry-level understanding of machine learning and statistics, and you program in Java, Python, or Scala, you’ll find these patterns useful for working on your own data applications.

Patterns include:

* Recommending music and the Audioscrobbler data set
* Predicting forest cover with decision trees
* Anomaly detection in network traffic with K-means clustering
* Understanding Wikipedia with Latent Semantic Analysis
* Analyzing co-occurrence networks with GraphX
* Geospatial and temporal data analysis on the New York City Taxi Trips data
* Estimating financial risk through Monte Carlo simulation
* Analyzing genomics data and the BDG project
* Analyzing neuroimaging data with PySpark and Thunder

# Learning Spark: Lightning-Fast Big Data Analysis

Data in all domains is getting bigger. How can you work with it efficiently? **Recently updated for Spark 1.3**, this book introduces Apache Spark, the open source cluster computing system that makes data analytics fast to write and fast to run. With Spark, you can tackle big datasets quickly through simple APIs in Python, Java, and Scala. This edition includes new information on Spark SQL, Spark Streaming, setup, and Maven coordinates.

Written by the developers of Spark, this book will have data scientists and engineers up and running in no time. You’ll learn how to express parallel jobs with just a few lines of code, and cover applications from simple batch jobs to stream processing and machine learning.

* Quickly dive into Spark capabilities such as distributed datasets, in-memory caching, and the interactive shell
* Leverage Spark’s powerful built-in libraries, including Spark SQL, Spark Streaming, and MLlib
* Use one programming paradigm instead of mixing and matching tools like Hive, Hadoop, Mahout, and Storm
* Learn how to deploy interactive, batch, and streaming applications
* Connect to data sources including HDFS, Hive, JSON, and S3
* Master advanced topics like data partitioning and shared variables

# Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale

Get ready to unlock the power of your data. With the fourth edition of this comprehensive guide, you’ll learn how to build and maintain reliable, scalable, distributed systems with Apache Hadoop. This book is ideal for programmers looking to analyze datasets of any size, and for administrators who want to set up and run Hadoop clusters.

Using Hadoop 2 exclusively, author Tom White presents new chapters on YARN and several Hadoop-related projects such as Parquet, Flume, Crunch, and Spark. You’ll learn about recent changes to Hadoop, and explore new case studies on Hadoop’s role in healthcare systems and genomics data processing.

* Learn fundamental components such as MapReduce, HDFS, and YARN
* Explore MapReduce in depth, including steps for developing applications with it
* Set up and maintain a Hadoop cluster running HDFS and MapReduce on YARN
* Learn two data formats: Avro for data serialization and Parquet for nested data
* Use data ingestion tools such as Flume (for streaming data) and Sqoop (for bulk data transfer)
* Understand how high-level data processing tools like Pig, Hive, Crunch, and Spark work with Hadoop
* Learn the HBase distributed database and the ZooKeeper distributed configuration service

**SQL book:**

<https://www.murachforinstructors.com/our-books-review-copies/all>

# Thanks for registering with us!

Thank you so much for taking the time to register with us! We're delighted that you're interested in reviewing our books, and we look forward to working with you.

We do, however, verify that you're an instructor before we fulfill your requests for review copies. We try to do that within 1 business day, and we will email you as soon as you have been verified.

Once you're verified, you can sign in at this website and then request the books, eBooks, and instructor's materials that you want to review for your courses.

If you have any questions or need more information, please [email us](mailto:judy@murach.com). Or call us at 1-800-221-5528 (in the US and Canada) or at 1-559-440-9071.

O’Reilly books:

## Thank you for your interest in O’Reilly books for use in your classroom

Did you know you can review O’Reilly books for use in your classroom with a 10-day free trial of the O’Reilly online learning platform? [Get your trial now.](https://learning.oreilly.com/register/) No credit card needed.

If you’ve already adopted the book for your classroom and would like to receive a digital desk copy, please email [adoption@oreilly.com](mailto:adoption@oreilly.com?subject=) with the following information:

* Title
* ISBN
* Institution
* Email address to deliver the ebook

To learn more about O’Reilly online learning for your academic institution, visit us at <https://www.oreilly.com/online-learning/academic.html>

If you have further questions about our Academic Solutions program or policies, please contact Customer Service at [adoption@oreilly.com](mailto:adoption@oreilly.com?subject=) or by phone[(707) 827-7019](tel:+1-707-827-7019) or [1-800-889-8969](tel:+1-800-889-8969).

# Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale 4th Edition

**ISBN-13:** 978-1491901632

**ISBN-10:** 1491901632