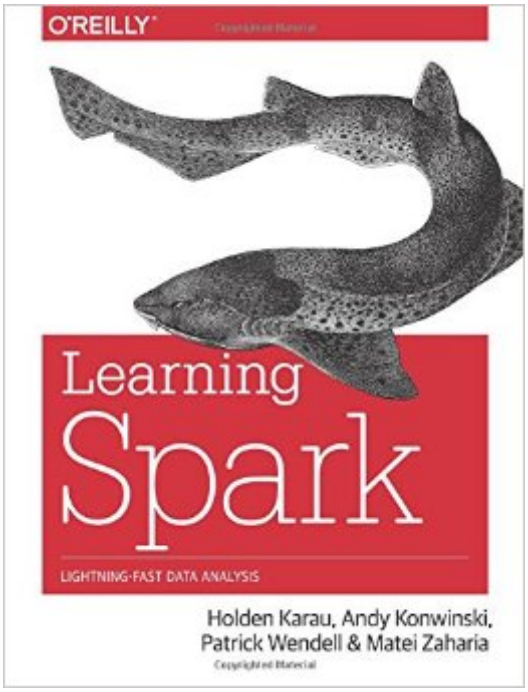
# Learning Spark: Lightning-Fast Big Data Analysis



## Synopsis

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. Youâ ™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 shellLeverage Sparkâ ™s powerful built-in libraries, including Spark SQL, Spark Streaming, and MLlibUse one programming paradigm instead of mixing and matching tools like Hive, Hadoop, Mahout, and StormLearn how to deploy interactive, batch, and streaming applicationsConnect to data sources including HDFS, Hive, JSON, and S3Master advanced topics like data partitioning and shared variables

## Book Information

Paperback: 276 pages

Publisher: O'Reilly Media; 1 edition (February 27, 2015)

Language: English

ISBN-10: 1449358624

ISBN-13: 978-1449358624

Product Dimensions: 7 x 0.6 x 9.2 inches

Shipping Weight: 1 pounds (View shipping rates and policies)

Average Customer Review: 4.0 out of 5 starsÂ Â See all reviewsÂ (51 customer reviews)

Best Sellers Rank: #10,304 in Books (See Top 100 in Books) #5 inÂ Books > Computers & Technology > Programming > Languages & Tools > Java #8 inÂ Books > Computers & Technology > Business Technology > Software > Enterprise Applications #8 inÂ Books > Computers & Technology > Databases & Big Data > Data Processing

## Customer Reviews

The textual components are well-written. However, the book tends to gloss over examples, providing the same, or similar, snippets to what's on the web site instead of providing full, working examples. This was a poor choice for a "learning" title. Instead of being able to work through each example, I found myself having to scroll around trying to figure out what was missing from each

snippet, and how to put together working code. Also, the book presents Scala, Java, and Python snippets throughout the early chapters, which is very distracting. I found myself having to mentally context-switch between the three languages instead of being able to following one all the way through. Would I buy it again? Maybe. There are just a few Spark books, and they're all pretty meh. You have to learn it somewhere, I guess.

I found this volume to be an excellent reference book for a Spark learner like me. I am a software developer, and several reviews suggested that this volume was too basic. I shouldn't have followed their advice. I bought an "advanced" book, instead, only to find myself left without material to fill in some important gaps. The information that is available on the Internet is great, but this book brings much of it together in one place. If you want to learn to think like a Spark programmer--*not* the same as thinking like a programmer--this is the place to begin.

I feel this is a decent compilation of the resources available over the internet. That way, it reduces the time needed for getting started with Spark. This book is definitely suitable anyone new to Spark and Big Data Processing. But for someone who has already worked with Spark and faced some challenges, this may not be helpful.

There are notions popped out out-of-the-blue throughout the book. One example is "accumulators". It is first seen in Chapter 3, while describing the aggregate function, yet lacking any definition of it. You have to go and check the programming guide on the official web page. The book is not comprehensive, but it is more like a companion.

This seems to be the definitive overview of Spark, and fortunately it's well written. It includes basics like setting up a Spark cluster, and topics like streaming and machine learning. I'm glad I read it.

Good text. I've purchased pretty much all of the other options out there-- while this book is still lacks detail (and borrows much of its content from the Databricks examples & Spark documentation available online) it's a worthwhile investment. The detailed treatment of RDDs, SparkSQL and streaming capability alone is enough to justify the cost.

The only reason for the 4-star rating and not higher is that the book is already a bit outdated (from a Scala perspective). Running newer versions of Spark do not support some of the examples in the

book. This does not change or distort the overall big picture of the book, however. Still a very intuitive and straight forward intro to Spark.

Disclaimer: I only write reviews when the item is a 5 star.This book is a very comprehensive introduction to Apache Spark. It covers srapk internals in detail. Armed with this knowledge, one can confidently advance to studying and implementing advanced machine learning application on Spark



Learning Spark: Lightning-Fast Big Data Analysis Machine Learning with Spark - Tackle Big Data with Powerful Spark Machine Learning Algorithms Analytics: Data Science, Data Analysis and Predictive Analytics for Business (Algorithms, Business Intelligence, Statistical Analysis, Decision Analysis, Business Analytics, Data Mining, Big Data) Data Analytics: What Every Business Must Know About Big Data And Data Science (Data Analytics for Business, Predictive Analysis, Big Data) Data Analytics: Practical Data Analysis and Statistical Guide to Transform and Evolve Any Business. Leveraging the Power of Data Analytics, Data ... (Hacking Freedom and Data Driven) (Volume 2) A collection of Advanced Data Science and Machine Learning Interview Questions Solved in Python and Spark (II): Hands-on Big Data and Machine ... Programming Interview Questions) (Volume 7) The Spark Story Bible: Spark a Journey through God's Word Lightning Fast Piano Scales: A Proven Method to Get Fast Piano Scales in 5 Minutes a Day (Piano Lessons, Piano Exercises) Unsupervised Machine Learning in Python: Master Data Science and Machine Learning with Cluster Analysis, Gaussian Mixture Models, and Principal Components Analysis Advanced Analytics with Spark: Patterns for Learning from Data at Scale Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results From Big Data to Big Profits: Success with Data and Analytics Unsupervised Deep Learning in Python: Master Data Science and Machine Learning with Modern Neural Networks written in Python and Theano (Machine Learning in Python) Deep Learning in Python Prerequisites: Master Data Science and Machine Learning with Linear Regression and Logistic Regression in Python (Machine Learning in Python) Convolutional Neural Networks in Python: Master Data Science and Machine Learning with Modern Deep Learning in Python, Theano, and TensorFlow (Machine Learning in Python) Deep Learning in Python: Master Data Science and Machine Learning with Modern Neural Networks written in Python, Theano, and TensorFlow (Machine Learning in Python) Lightning-fast Spanish for Kids and Families: Learn Spanish, Speak Spanish, Teach Kids Spanish - Quick as a Flash, Even if You Don't Speak a Word Now! (Spanish Edition) Lightning-Fast French for Kids and Families: Learn

French, Speak French, Teach Kids French - Quick as a Flash, Even if You Don’t Speak a Word Now! Credit Repair Ninja (A 5 Minute Guide) - 21 Ways To Fix Your Credit Score Lightning Fast - 2016: How To Fix Your Bad Credit Score In 30 Days Or Less