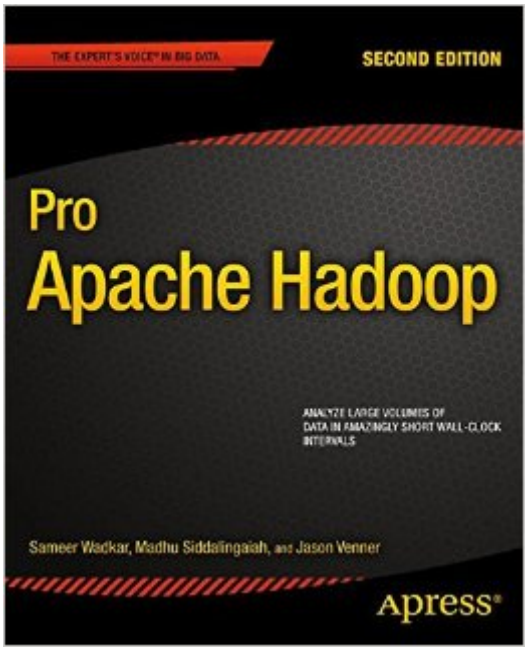
# Pro Apache Hadoop



## Synopsis

Pro Apache Hadoop, Second Edition brings you up to speed on Hadoop â " the framework of big data. Revised to cover Hadoop 2.0, the book covers the very latest developments such as YARN (aka MapReduce 2.0), new HDFS high-availability features, and increased scalability in the form of HDFS Federations. All the old content has been revised too, giving the latest on the ins and outs of MapReduce, cluster design, the Hadoop Distributed File System, and more. This book covers everything you need to build your first Hadoop cluster and begin analyzing and deriving value from your business and scientific data. Learn to solve big-data problems the MapReduce way, by breaking a big problem into chunks and creating small-scale solutions that can be flung across thousands upon thousands of nodes to analyze large data volumes in a short amount of wall-clock time. Learn how to let Hadoop take care of distributing and parallelizing your softwareâ •you just focus on the code; Hadoop takes care of the rest. Covers all that is new in Hadoop 2.0  Written by a professional involved in Hadoop since day one  Takes you quickly to the seasoned pro level on the hottest cloud-computing frameworkÂ Â

## Book Information

Series: Pro

Paperback: 444 pages

Publisher: Apress; 2nd ed. edition (September 10, 2014)

Language: English

ISBN-10: 1430248637

ISBN-13: 978-1430248637

Product Dimensions:  7.5 x 1 x 9.2 inches

Shipping Weight: 2.1 pounds (View shipping rates and policies)

Average Customer Review:  4.7 out of 5 starsÂ Â See all reviewsÂ (16 customer reviews)

Best Sellers Rank: #830,500 in Books (See Top 100 in Books)  #81 inÂ Books > Computers & Technology > Programming > Parallel Programming  #470 inÂ Books > Computers & Technology > Databases & Big Data > Data Mining  #916 inÂ Books > Textbooks > Computer Science > Database Storage & Design

## Customer Reviews

This book is excellent, and this is coming from a guy who hated almost every other hadoop book he read. There are a couple things that these authors do different from other hadoop authors:1. They thoroughly explain the difference between all the hadoop versions, and what things are

compatible/incompatible with one another. It covers YARN in detail.2. They explain how to do things using both the old and the new hadoop APIs. This is so important because a lot of code has already been developed for the old API, and people are likely to encounter that at work.3. They include import statements in their code samples, I can't tell you how many times I read other documentation that omitted the import statements and it made things so confusing for me. This is especially important with hadoop, because there are a lot of classes in their APIs with the same names but in different packages.4. They have a hadoop administration chapter, which is a big deal to me because I have struggled desperately to make hadoop run in the past. It is not easy by any means. They cover multiple ways to get hadoop running: using cloudera's VM, using 's cloud, and setting things up yourself. They explain the config files and what all of their settings do too.I would recommend reading this book BEFORE you read "Hadoop: The Definitive Guide" because this will get you up to speed faster. Since the definitive guide gives you a lot of fine details, it is not a good starter book.

As another reviewer mentioned, this is an excellent first book for Hadoop. It covers Hadoop 2.0 in great detail while still staying within familiar ground for most readers since it uses SQL to explain Hadoop concepts. The code examples are consistent and help illustrate the points. The chapters on Hadoop administration focus on the practical aspects you are likely to use at work.One area where the book shines is in its treatment of Large Table Joins. Secondary Sort is an important and extremely practical concept which is usually only mentioned in passing in other books. But this book devotes almost an entire chapter to it and demonstrates its use for Large Table Joins. The authors have gone deep into the Hadoop source code and described the underlying Hadoop framework where appropriate.The treatment of Hadoop libraries like HBase and Pig is also unique. Instead of being focused on just the usage of these libraries the focus is on explaining the underlying concepts and the describing how these libraries work internally as well as where their use is appropriate.The book leave you with a sense of not just how Hadoop should be used but also how Hadoop works on the inside. A great effort from the authors.

If you're new to Hadoop and especially to programming with MapReduce, this is a great book to get started. There is a lot of attention paid to details and I like the presentation of SQL clauses as MapReduce constructs.I'm working with two of the authors on a big data project that handles billions of transactions per day and their experience shines through, especially in the chapters that deal with MapReduce. The use of secondary sorting, partitioning and other useful techniques are covered

thoroughly. The authors have also spent a lot of time looking at the source code and provide useful insights that are not part of the documentation.The editors could have done a better job of proofreading, but it doesn't detract from the outstanding content.

Im already familiar with MRv1 but YARN was little bit confusing for me.A lot of IT books are full of dummy stories (Ex-wife, kids vacation, etc...) things that the reader doesn't care about and most the time ittakes the reader outside out the book BUT this books isn't like that at all.So far I read only the first 3 chapters and it's amazing !!!It really help me fully understand HDFS architecture and YARN (Resource manager, node manager, etc...)

This was my first book about Hadoop and I found the explanations and techniques to be well written. The authors are experts in the field as they point out many gotchas along the way. Knowing some Java is a prerequisite for this book as there are a lot of code examples.Most of the first half of the book covers Hadoop directly including how to configure Hadoop, how to do MapReduce, and other common techniques. The later part of the book looks a lot at tools that interact with Hadoop such as Pig, Hive, HCatalog, HBase, Hama, Spark and some others.The main example used in this book is word count spread across many documents. This example worked well as it is was easy to understand, and it was the same topic that the original MapReduce was based on. Over the course of the book the authors allow the example to get more and more complex, while providing clear explanations.I found the flow and the writing of the book to be excellent. At the beginning there are clear high level explanations of different Big Data tools. There is no fluff at all, the book is purely technical in this sense.

This is a great guide for the practical use of Hadoop 2 with just the right amount of breadth and depth. The book tells you what is new with Hadoop and then gives you examples on how to use these new features. From the examples, you can also see that the authors are actual Hadoop implementers and not just professional tech authors. The coverage of Hadoop "sweet spots" like data warehousing, ETL and data analysis, will let you hit the ground running with your Hadoop implementation instead of having to reinvent the wheel each time.

Not much to post, I normally evaluate something like this by deciding if I would or wouldn't buy it again; in other words, a five or a one. Strong points included a focus on newer technologies such as YARN, and special kudos for using SQL as the example for explaining map reduce. I would buy it

again.

I’ve had the privilege of working on some projects with two of the authors and to witness the level of attention and time they poured into this book. I have a copy and am looking forward to learning from it.