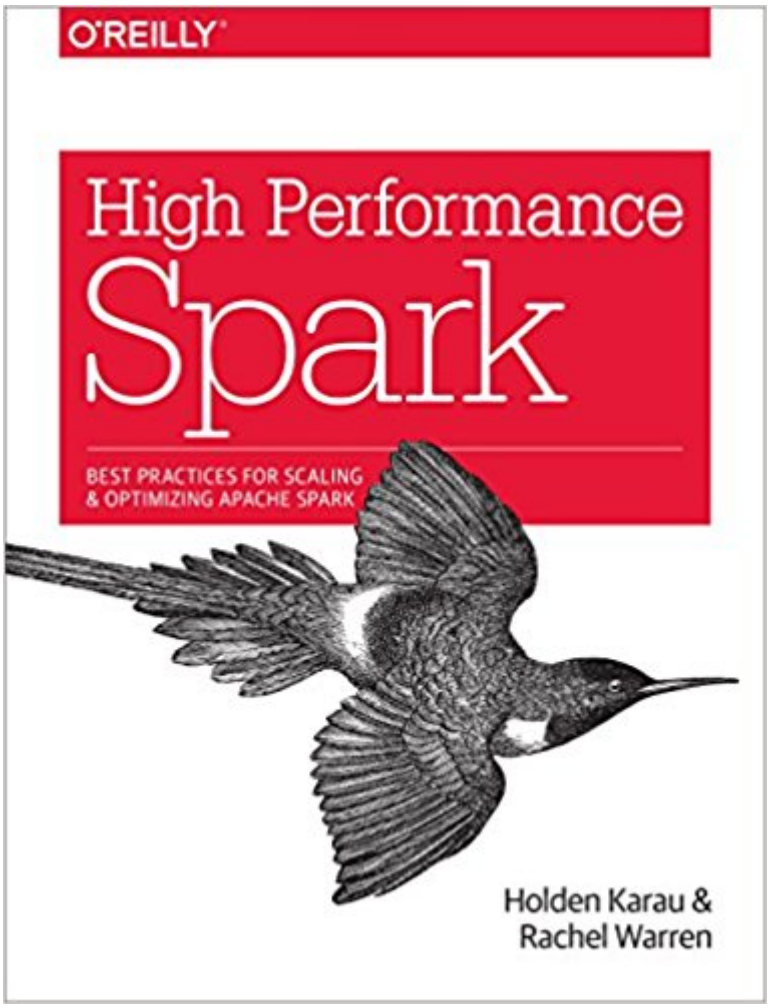
High Performance Spark: Best Practices For Scaling And Optimizing Apache Spark



Synopsis

Apache Spark is amazing when everything clicks. But if you havenÃ¢â ¬â„¢t seen the performance improvements you expected, or still donÃ¢â ¬â„¢t feel confident enough to use Spark in production, this practical book is for you. Authors Holden Karau and Rachel Warren demonstrate performance optimizations to help your Spark queries run faster and handle larger data sizes, while using fewer resources.Ideal for software engineers, data engineers, developers, and system administrators working with large-scale data applications, this book describes techniques that can reduce data infrastructure costs and developer hours. Not only will you gain a more comprehensive understanding of Spark, youÃ¢â ¬â„¢ll also learn how to make it sing.With this book, youÃ¢â ¬â„¢ll explore:How Spark SQLÃ¢â ¬â„¢s new interfaces improve performance over SQLÃ¢â ¬â„¢s RDD data structureThe choice between data joins in Core Spark and Spark SQLTechniques for getting the most out of standard RDD transformationsHow to work around performance issues in SparkÃ¢â ¬â„¢s key/value pair paradigmWriting high-performance Spark code without Scala or the JVMHow to test for functionality and performance when applying suggested improvementsUsing Spark MLlib and Spark ML machine learning librariesSparkÃ¢â ¬â„¢s Streaming components and external community packages

Book Information

Paperback: 358 pages

Publisher: O'Reilly Media; 1 edition (June 16, 2017)

Language: English

ISBN-10: 1491943203

ISBN-13: 978-1491943205

Product Dimensions: 7 x 0.7 x 9.2 inches

Shipping Weight: 1.8 pounds (View shipping rates and policies)

Average Customer Review: 4.7 out of 5 stars 7 customer reviews

Best Sellers Rank: #23,717 in Books (See Top 100 in Books) #4 inÃ Â Books > Computers & Technology > Databases & Big Data > Data Warehousing #5 inÃ Â Books > Computers & Technology > Software > Databases #21 inÃ Â Books > Computers & Technology > Databases & Big Data > Data Mining

Customer Reviews

Best practices for scaling and optimizing Apache Spark

Holden Karau is transgender Canadian, and an active open source contributor. When not in San Francisco working as a software development engineer at IBM's Spark Technology Center, Holden talks internationally on Apache Spark and holds office hours at coffee shops at home and abroad. She is a Spark committer with frequent contributions, specializing in PySpark and Machine Learning. Prior to IBM she worked on a variety of distributed, search, and classification problems at Alpine, Databricks, Google, Foursquare, and . She graduated from the University of Waterloo with a Bachelor of Mathematics in Computer Science. Outside of software she enjoys playing with fire, welding, scooters, poutine, and dancing.Rachel Warren is a data scientist and software engineer at Alpine Data Labs, where she uses Spark to address real world data processing challenges. She has experience working as an analyst both in industry and academia. She graduated with a degree in Computer Science from Wesleyan University in Connecticut.

Overall, I thought this was a very good book. It strikes a good balance between detailed instruction and depth and being a guidebook, not an instruction manual. It's the usual high quality that I've come to expect from O'Reilly, and I feel much more confident about my understanding of Spark, both as a user and of the inner workings.Much of the book is written with a focus on performance. There's some discussion of statistical concepts, but the book is clearly aimed at helping the reader use Spark in a resource-efficient manner (which makes a lot of sense, given that Spark comes into play when you're tackling large data sets).Virtually all of the code examples are written in Scala. When I began reading, my Scala abilities were fairly limited, but the authors do a good job of parsing and commenting on the code such that I now feel much stronger in Scala, as well. They do have a chapter that discusses using Python and Java (including JVM), but most of the book is presented through Scala.My one complaint about this book is that it's a bit heavy on the code. It's possible that it's necessary, but I ended up skimming most of the coding examples, and it made for some tedious reading at times. Then again, there were several examples that I scrutinized closely, and having thorough examples did help me learn quite a bit of Scala.

Apache spark moves along the continuum of parallel processing. ÃƒÂ¢Ã â ¬Ã Å“Apache Spark is a high-performance, general-purpose distributed computing system that has become the most active Apache open source project with more than 1,000 active contributorsÃƒÂ¢Ã â ¬Ã Â•. The authors go on to state ÃƒÂ¢Ã â ¬Ã Å“Spark enables us to process large quantities of data, beyond what can fit on a single machine, with a high-level, relatively easy-to-use APIÃƒÂ¢Ã â ¬Ã Â•. Most people in (and out) of IT will never have any contact with Spark. I need to know about it only

because my job involves having at least a superficial knowledge of every significant aspect of IT. This book presumes you are already conversant with Apache Spark and need no education or hand-holding in that regard. Rather this bookÃƒÂ¢Ã â ¬Ã â„¢s goal is to help the reader make their Spark queries Ã ƒÂ¢Ã â ¬Ã Å“faster, able to handle larger data sizes, and use fewer resourcesÃƒÂ¢Ã â ¬Ã Â•. Being able to at least read Scala is highly recommended. The entire book is loaded with detailed examples. For the casual reader, such as myself, lacking a Spark environment to play in, there is an empty feeling Ã ƒÂ¢Ã â ¬Ã â œ you can read the examples, study them, but not run them. Having read literally dozens or more programming cookbooks during the course of my career, this one feels right, but without being able to run the examples, thatÃƒÂ¢Ã â ¬Ã â„¢s just as an assumption. It does, however, make me wish I had some huge datasets to work on. Maybe I can get a job with the NSA? I bet there are a lot of Spark experts there.Jerry

much of the other learning material on spark is something like "if you want to join two RDDs you can use the following function"this book however dedicates several chapters explaining it in detail and making the reader understand the internals and the performance implications.

This book is heavily Scala centric and for beginners the only takeaway should be that you should be fairly comfortable with Scala If you hope to have a "Spark" Centered carrear. If you are in the big data / Warehouse space with Spark in the center of action, I highly recommend this book. It focuses heavily on all areas of Performance. You can keep this book handy as a reference guide as well.Good job.

I’m new to Spark (but not new to the space), and I found this book to be great. It doesn’t assume you’re a total beginner, but it was easy enough to fill those gaps with a little online research.I prefer that the book skips past the basics and focuses on more advanced examples. If you’re self sufficient like that, then this book is fine even if you’re a beginner.

This book clarifies lots of my questions on Spark. I especially appreciate the walk through joins.

Worth reading.

High Performance Spark: Best Practices for Scaling and Optimizing Apache Spark High Fiber

Recipes: 101 Quick and Easy High Fiber Recipes for Breakfast, Snacks, Side Dishes, Dinner and Dessert (high fiber cookbook, high fiber diet, high fiber recipes, high fiber cooking) Scaling and Integration of High-Speed Electronics and Optomechanical Systems (Selected Topics in Electronics and Systems) The Spark Story Bible: Spark a Journey through God's Word Becoming a Supple Leopard 2nd Edition: The Ultimate Guide to Resolving Pain, Preventing Injury, and Optimizing Athletic Performance Optimizing Jet Transport Efficiency: Performance, Operations, and Economics Neurological Rehabilitation: Optimizing motor performance, 2e Portraits of "The Whiteman": Linguistic Play and Cultural Symbols Among the Western Apache Wisdom Sits in Places: Landscape and Language Among the Western Apache Victorio: Apache Warrior and Chief (The Oklahoma Western Biographies) Apache Warrior vs US Cavalryman: 1846Ã¢â ¬â œ86 (Combat) Indeh: A Story of the Apache Wars In the Great Apache Forest (1920) High Blood Pressure Cure: How To Lower Blood Pressure Naturally in 30 Days (Alternative Medicine, Natural Cures, Natural Remedies, High Blood Pressure ... Cures for High Blood Pressure, High Bl) Scaling Up: How a Few Companies Make It...and Why the Rest Don't (Rockefeller Habits 2.0) Innovating: A Doer's Manifesto for Starting from a Hunch, Prototyping Problems, Scaling Up, and Learning to Be Productively Wrong Agile and Lean Program Management: Scaling Collaboration Across the Organization Innovating: A Doer's Manifesto for Starting from a Hunch, Prototyping Problems, Scaling Up, and Learning to Be Productively Wrong (MIT Press) Fractals and Scaling in Finance: Discontinuity, Concentration, Risk. Selecta Volume E Root Scaling and Planing: A Fundamental Therapy