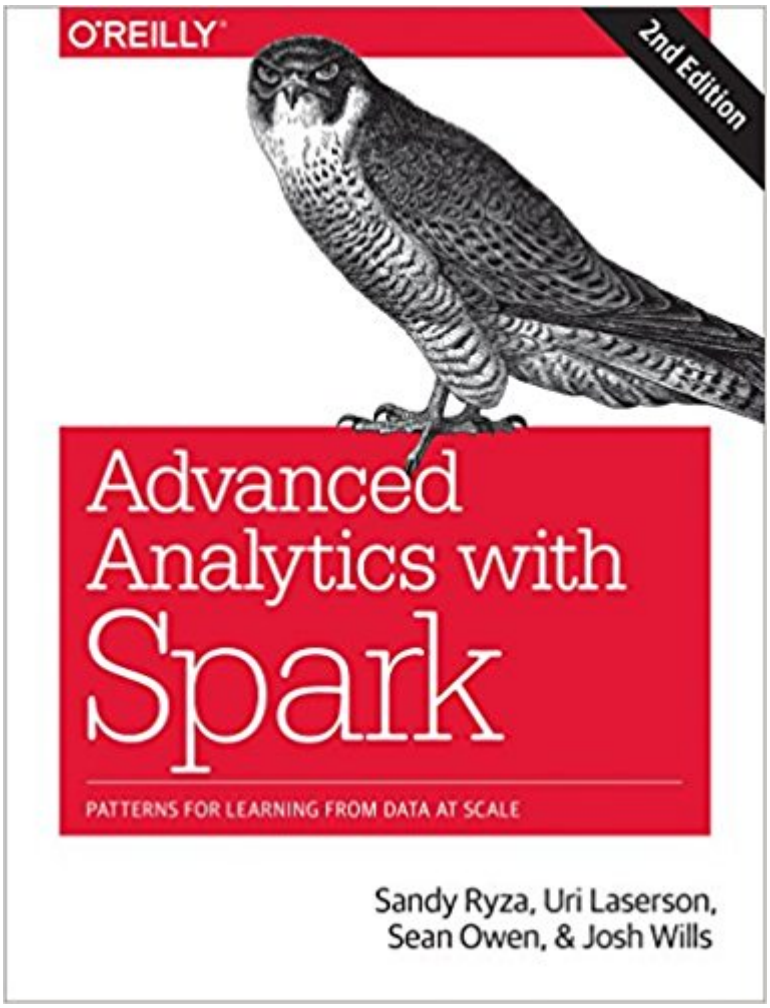
Advanced Analytics With Spark: Patterns For Learning From Data At Scale



Synopsis

In the second edition of 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. Updated for Spark 2.1, this edition acts as an introduction to these techniques and other best practices in Spark programming.YouÃ¢â ¬â„¢ll start with an introduction to Spark and its ecosystem, and then dive into patterns that apply common techniquesÃ¢â ¬â •including classification, clustering, collaborative filtering, and anomaly detectionÃ¢â ¬â •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, youÃ¢â ¬â„¢ll find the bookÃ¢â ¬â„¢s patterns useful for working on your own data applications.With this book, you will:Familiarize yourself with the Spark programming modelBecome comfortable within the Spark ecosystemLearn general approaches in data scienceExamine complete implementations that analyze large public data setsDiscover which machine learning tools make sense for particular problemsAcquire code that can be adapted to many uses

Book Information

Paperback: 280 pages

Publisher: O'Reilly Media; 2 edition (July 6, 2017)

Language: English

ISBN-10: 1491972955

ISBN-13: 978-1491972953

Product Dimensions: 7 x 0.5 x 9.1 inches

Shipping Weight: 1.2 pounds (View shipping rates and policies)

Average Customer Review: Be the first to review this item

Best Sellers Rank: #27,920 in Books (See Top 100 in Books) #10 inÃ Â Books > Textbooks > Computer Science > Algorithms #18 inÃ Â Books > Computers & Technology > Programming > Algorithms #23 inÃ Â Books > Computers & Technology > Databases & Big Data > Data Mining

Customer Reviews

View larger From the Preface WhatÃƒÂ¢Ã â ¬Ã â„¢s in This Book The first chapter will place Spark within the wider context of data science and big data analytics. After that, each chapter will comprise a self-contained analysis using Spark. The second chapter will introduce the basics of data processing in Spark and Scala through a use case in data cleansing. The next few chapters

will delve into the meat and potatoes of machine learning with Spark, applying some of the most common algorithms in canonical applications. The remaining chapters are a bit more of a grab bag and apply Spark in slightly more exotic applicationsÃƒÂ¢Ã â ¬Ã â •for example, querying Wikipedia through latent semantic relationships in the text or analyzing genomics data. The Second Edition Since the first edition, Spark has experienced a major version upgrade that instated an entirely new core API and sweeping changes in subcomponents like MLlib and Spark SQL. In the second edition, weÃƒÂ¢Ã â ¬Ã â„¢ve made major renovations to the example code and brought the materials up to date with SparkÃƒÂ¢Ã â ¬Ã â„¢s new best practices.

Sandy Ryza develops algorithms for public transit at Remix. Prior, he was a senior data scientist at Cloudera and Clover Health. He is an Apache Spark committer, Apache Hadoop PMC member, and founder of the Time Series for Spark project. He holds the Brown University computer science department's 2012 Twining award for "Most Chill".Uri Laserson is an Assistant Professor of Genetics at the Icahn School of Medicine at Mount Sinai, where he develops scalable technology for genomics and immunology using the Hadoop ecosystem.Sean Owen is Director of Data Science at Cloudera. He is an ApacheSpark committer and PMC member, and was an Apache Mahout committer.Josh Wills is the Head of Data Engineering at Slack, the founder of the Apache Crunch project, and wrote a tweet about data scientists once.

Book Series) Analytics: Data Science, Data Analysis and Predictive Analytics for Business R for Everyone: Advanced Analytics and Graphics (Addison-Wesley Data and Analytics) R for Everyone: Advanced Analytics and Graphics (2nd Edition) (Addison-Wesley Data & Analytics Series) Cutting Edge Marketing Analytics: Real World Cases and Data Sets for Hands On Learning (FT Press Analytics) Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data The Analytics Revolution: How to Improve Your Business By Making Analytics Operational In The Big Data Era The Spark Story Bible: Spark a Journey through God's Word High Performance Spark: Best Practices for Scaling and Optimizing Apache Spark Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results Data Management: Foundations of Data Analytics Data Analytics for Beginners: Your Ultimate Guide to Learn and Master Data Analysis Unsupervised Machine Learning in Python: How to Find Distinct Patterns in Your Data Without Being at the Mercy of Data Labeling by Third-Party Workers