
Big Data Analytics With Spark

Home Springer

Advanced Analytics with Pyspark

Introduction to Hadoop, Spark, and Machine-Learning

Designing and Building Effective Analytics at Scale

Get Started With Big Data Analytics Using Apache Spark

Big Data Analytics Beyond Hadoop

Spark for Data Science

Data Analytics with Spark Using Python

Learning Spark

Practical Data Science with Hadoop and Spark

Learning Spark

Lightning-Fast Big Data Analysis

A Practitioner's Guide to Using Spark for Large Scale Data Analysis

Next-Generation Big Data

Build highly effective analytics solutions to gain valuable insight into your big data

Learning PySpark

A beginner's guide to harnessing the power and ease of PySpark 3

Mastering Spark with R

Patterns for Learning from Data at Scale

Big Data Analytics with Hadoop 3

Big Data Analytics Beyond Hadoop

The Complete Guide to Large-Scale Analysis and Modeling

BIG DATA ANALYTICS

Big Data Analytics with Java

Hands-On Big Data Analytics with PySpark

Practical Big Data Analytics

Big Data Analytics with Spark

Essential PySpark for Scalable Data Analytics

Big Data Processing Using Spark in Cloud

Explore the concepts of functional programming, data streaming, and machine learning

Spark: The Definitive Guide

Machine Learning in Python

Combine Spark and Python to Unlock the Powers of Parallel Computing and Machine Learning

Scala Programming for Big Data Analytics

Big Data Analytics with Spark and Hadoop

Big Data Analytics with Spark

Big Data Processing Made Simple

Big Data Analytics

Advanced Analytics with Spark

Hands-on techniques to implement enterprise analytics and machine learning using Hadoop, Spark, NoSQL and R

Big Data Analytics With Spark Home Springer
Downloaded from archive.imba.com
by guest

ALEXANDER BRIGHT

Advanced Analytics with Pyspark Packt Publishing Ltd

Get command of your organizational Big Data using the power of data science and analytics Key Features A perfect companion to boost your Big Data storing, processing, analyzing skills to help you take informed business decisions Work with the best tools such as Apache Hadoop, R, Python, and Spark for NoSQL platforms to perform massive online analyses Get expert tips on statistical inference, machine learning, mathematical modeling, and data visualization for Big Data Book Description Big Data analytics relates to the strategies used by organizations to collect, organize and analyze large amounts of data to uncover valuable business insights that otherwise cannot be analyzed through traditional systems. Crafting an enterprise-scale cost-efficient Big Data and machine learning solution to uncover insights and

value from your organization's data is a challenge. Today, with hundreds of new Big Data systems, machine learning packages and BI Tools, selecting the right combination of technologies is an even greater challenge. This book will help you do that. With the help of this guide, you will be able to bridge the gap between the theoretical world of technology with the practical ground reality of building corporate Big Data and data science platforms. You will get hands-on exposure to Hadoop and Spark, build machine learning dashboards using R and R Shiny, create web-based apps using NoSQL databases such as MongoDB and even learn how to write R code for neural networks. By the end of the book, you will have a very clear and concrete understanding of what Big Data analytics means, how it drives revenues for organizations, and how you can develop your own Big Data analytics solution using different tools and methods articulated in this book. What you will learn - Get a

360-degree view into the world of Big Data, data science and machine learning - Broad range of technical and business Big Data analytics topics that caters to the interests of the technical experts as well as corporate IT executives - Get hands-on experience with industry-standard Big Data and machine learning tools such as Hadoop, Spark, MongoDB, KDB+ and R - Create production-grade machine learning BI Dashboards using R and R Shiny with step-by-step instructions - Learn how to combine open-source Big Data, machine learning and BI Tools to create low-cost business analytics applications - Understand corporate strategies for successful Big Data and data science projects - Go beyond general-purpose analytics to develop cutting-edge Big Data applications using emerging technologies Who this book is for The book is intended for existing and aspiring Big Data professionals who wish to become the go-to person in their organization when it comes to Big Data architecture, analytics, and governance. While no

prior knowledge of Big Data or related technologies is assumed, it will be helpful to have some programming experience.

[Introduction to Hadoop, Spark, and Machine-Learning](#) "O'Reilly Media, Inc."

The book describes the emergence of big data technologies and the role of Spark in the entire big data stack. It compares Spark and Hadoop and identifies the shortcomings of Hadoop that have been overcome by Spark. The book mainly focuses on the in-depth architecture of Spark and our understanding of Spark RDDs and how RDD complements big data's immutable nature, and solves it with lazy evaluation, cacheable and type inference. It also addresses advanced topics in Spark, starting with the basics of Scala and the core Spark framework, and exploring Spark data frames, machine learning using Mllib, graph analytics using Graph X and real-time processing with Apache Kafka, AWS Kinesis, and Azure Event Hub. It then goes on to investigate Spark using PySpark and R. Focusing on the current big data stack, the book examines

the interaction with current big data tools, with Spark being the core processing layer for all types of data. The book is intended for data engineers and scientists working on massive datasets and big data technologies in the cloud. In addition to industry professionals, it is helpful for aspiring data processing professionals and students working in big data processing and cloud computing environments.

Designing and Building Effective Analytics at Scale "O'Reilly Media, Inc."

Gain the key language concepts and programming techniques of Scala in the context of big data analytics and Apache Spark. The book begins by introducing you to Scala and establishes a firm contextual understanding of why you should learn this language, how it stands in comparison to Java, and how Scala is related to Apache Spark for big data analytics. Next, you'll set up the Scala environment ready for examining your first Scala programs. This is followed by sections on Scala fundamentals including mutable/immutable variables, the type

hierarchy system, control flow expressions and code blocks. The author discusses functions at length and highlights a number of associated concepts such as functional programming and anonymous functions. The book then delves deeper into Scala's powerful collections system because many of Apache Spark's APIs bear a strong resemblance to Scala collections. Along the way you'll see the development life cycle of a Scala program. This involves compiling and building programs using the industry-standard Scala Build Tool (SBT). You'll cover guidelines related to dependency management using SBT as this is critical for building large Apache Spark applications. Scala Programming for Big Data Analytics concludes by demonstrating how you can make use of the concepts to write programs that run on the Apache Spark framework. These programs will provide distributed and parallel computing, which is critical for big data analytics. What You Will Learn See the fundamentals of Scala as a general-purpose programming language Understand functional

programming and object-oriented programming constructs in Scala Use Scala collections and functions Develop, package and run Apache Spark applications for big data analytics Who This Book Is For Data scientists, data analysts and data engineers who intend to use Apache Spark for large-scale analytics. /div
Get Started With Big Data Analytics Using Apache Spark Packt Publishing Ltd
 Solve Data Analytics Problems with Spark, PySpark, and Related Open Source Tools Spark is at the heart of today's Big Data revolution, helping data professionals supercharge efficiency and performance in a wide range of data processing and analytics tasks. In this guide, Big Data expert Jeffrey Aven covers all you need to know to leverage Spark, together with its extensions, subprojects, and wider ecosystem. Aven combines a language-agnostic introduction to foundational Spark concepts with extensive programming examples utilizing the popular and intuitive PySpark development environment. This guide's focus on Python makes it

widely accessible to large audiences of data professionals, analysts, and developers—even those with little Hadoop or Spark experience. Aven's broad coverage ranges from basic to advanced Spark programming, and Spark SQL to machine learning. You'll learn how to efficiently manage all forms of data with Spark: streaming, structured, semi-structured, and unstructured. Throughout, concise topic overviews quickly get you up to speed, and extensive hands-on exercises prepare you to solve real problems. Coverage includes:

- Understand Spark's evolving role in the Big Data and Hadoop ecosystems
- Create Spark clusters using various deployment modes
- Control and optimize the operation of Spark clusters and applications
- Master Spark Core RDD API programming techniques
- Extend, accelerate, and optimize Spark routines with advanced API platform constructs, including shared variables, RDD storage, and partitioning
- Efficiently integrate Spark with both SQL and nonrelational data stores
- Perform stream processing and messaging

with Spark Streaming and Apache Kafka • Implement predictive modeling with SparkR and Spark MLlib
Big Data Analytics Beyond Hadoop Packt Publishing Ltd
 This book introduces Apache Spark, the open source cluster computing system that makes data analytics fast to write and fast to run. 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.--
Spark for Data Science Apress
 Ready to use statistical and machine-learning techniques across large data sets? This practical guide shows you why the Hadoop ecosystem is perfect for the job. Instead of deployment, operations, or software development usually associated with distributed computing, you'll focus on particular analyses you can build, the data warehousing techniques that Hadoop provides, and higher order data workflows this framework can produce. Data scientists and analysts will learn how to perform a wide range of techniques, from writing

MapReduce and Spark applications with Python to using advanced modeling and data management with Spark MLlib, Hive, and HBase. You'll also learn about the analytical processes and data systems available to build and empower data products that can handle—and actually require—huge amounts of data. Understand core concepts behind Hadoop and cluster computing Use design patterns and parallel analytical algorithms to create distributed data analysis jobs Learn about data management, mining, and warehousing in a distributed context using Apache Hive and HBase Use Sqoop and Apache Flume to ingest data from relational databases Program complex Hadoop and Spark applications with Apache Pig and Spark DataFrames Perform machine learning techniques such as classification, clustering, and collaborative filtering with Spark's MLlib

Data Analytics with Spark Using Python
Packt Publishing Ltd
Utilize this practical and easy-to-follow guide to modernize traditional enterprise data warehouse and business intelligence environments

with next-generation big data technologies. Next-Generation Big Data takes a holistic approach, covering the most important aspects of modern enterprise big data. The book covers not only the main technology stack but also the next-generation tools and applications used for big data warehousing, data warehouse optimization, real-time and batch data ingestion and processing, real-time data visualization, big data governance, data wrangling, big data cloud deployments, and distributed in-memory big data computing. Finally, the book has an extensive and detailed coverage of big data case studies from Navistar, Cerner, British Telecom, Shopzilla, Thomson Reuters, and Mastercard. What You'll Learn Install Apache Kudu, Impala, and Spark to modernize enterprise data warehouse and business intelligence environments, complete with real-world, easy-to-follow examples, and practical advice Integrate HBase, Solr, Oracle, SQL Server, MySQL, Flume, Kafka, HDFS, and Amazon S3 with Apache Kudu, Impala, and Spark Use StreamSets, Talend, Pentaho, and CDAP for

real-time and batch data ingestion and processing Utilize Trifacta, Alteryx, and Datameer for data wrangling and interactive data processing Turbocharge Spark with Alluxio, a distributed in-memory storage platform Deploy big data in the cloud using Cloudera Director Perform real-time data visualization and time series analysis using Zoomdata, Apache Kudu, Impala, and Spark Understand enterprise big data topics such as big data governance, metadata management, data lineage, impact analysis, and policy enforcement, and how to use Cloudera Navigator to perform common data governance tasks Implement big data use cases such as big data warehousing, data warehouse optimization, Internet of Things, real-time data ingestion and analytics, complex event processing, and scalable predictive modeling Study real-world big data case studies from innovative companies, including Navistar, Cerner, British Telecom, Shopzilla, Thomson Reuters, and Mastercard Who This Book Is For BI and big data warehouse professionals interested in gaining practical and real-world

insight into next-generation big data processing and analytics using Apache Kudu, Impala, and Spark; and those who want to learn more about other advanced enterprise topics

[Learning Spark](#) O'Reilly Media

A handy reference guide for data analysts and data scientists to fetch "Value" out of big data analytics using Spark on Hadoop Clusters About This Book* Practical tutorial with real-world examples that explores Spark on Hadoop clusters* This book is based on the latest version of Apache Spark and Hadoop integrated with the most commonly used tools* Learn about all the Spark stack components including the latest topics such as DataFrames, DataSets, and SparkR Who This Book Is For Though this book is primarily aimed at data analysts and data scientists, it will also help architects, programmers, and practitioners.

Knowledge of either Spark or Hadoop would be beneficial. It is assumed that you have basic programming background in Scala, Python, SQL, or R programming with basic Linux experience. Working experience within big data

environments is not mandatory. What You Will Learn* Find out about and implement the tools and techniques of big data analytics using Spark on Hadoop clusters* Understand all the Hadoop and Spark ecosystem components and how Spark replaced MapReduce* Get to know all the Spark components: Spark Core, Spark SQL, DataFrames, DataSets, Streaming, MLlib, and Graphx* See batch and real-time data analytics using Spark Core, Spark SQL, and Spark Streaming* Get to grips with data science and machine learning using MLlib, H2O, Hivemall, Graphx, and SparkR* Get an introduction to all the new tools (based on Notebooks, Data Flow, and Spark as a Service) and their integrations with Spark and Hadoop In Detail This book explains the fundamentals of Apache Spark and Hadoop, and how they are easily integrated together with the most commonly used tools and techniques. All the Spark components-Spark Core, Spark SQL, DataFrames, Data sets, Streaming, MLlib, Graphx, and Hadoop core components-HDFS, MapReduce, and Yarn are explored in

greater depth with implementation examples on Spark and Hadoop clusters. The big data analytics industry is moving away from MapReduce to Spark. In this book, the advantages of Spark over MapReduce are explained at great depth so you can reap the benefits of in-memory speeds. The DataFrames API, Data Sources API, and new Data sets API are explained so you can build big data analytical applications. We'll explore real-time data analytics using Spark Streaming with Apache Kafka and HBase to help you build streaming applications. You'll get to know the machine learning techniques using MLlib and SparkR, and Graph Analytics with the GraphX component of Spark. You will also get the opportunity to start working with web-based notebooks such as Jupyter, Apache Zeppelin, and the data flow tool Apache NiFi to analyze and visualize data. [Practical Data Science with Hadoop and Spark](#) Packt Publishing Ltd 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 Packt Publishing Ltd

In this book, you'll learn to implement some practical and proven techniques to improve aspects of programming and administration in Apache Spark. Techniques are demonstrated using practical examples and best practices. You will also learn how to use Spark and its Python API to create performant analytics with large-scale data.

[Lightning-Fast Big Data Analysis](#) Lulu.com

If you're like most R users, you have deep knowledge and love for statistics. But as your organization continues to collect huge amounts of data, adding tools such as Apache Spark makes a lot of sense. With this practical book, data scientists and professionals working with large-scale data applications will learn how to use Spark from R to tackle big data and big compute problems. Authors Javier Luraschi, Kevin Kuo, and Edgar Ruiz show you how to use R with Spark to solve different data analysis problems. This book covers relevant data science topics, cluster

computing, and issues that should interest even the most advanced users. Analyze, explore, transform, and visualize data in Apache Spark with R Create statistical models to extract information and predict outcomes; automate the process in production-ready workflows Perform analysis and modeling across many machines using distributed computing techniques Use large-scale data from multiple sources and different formats with ease from within Spark Learn about alternative modeling frameworks for graph processing, geospatial analysis, and genomics at scale Dive into advanced topics including custom transformations, real-time data processing, and creating custom Spark extensions *A Practitioner's Guide to Using Spark for Large Scale Data Analysis* "O'Reilly Media, Inc." Frank Kane's hands-on Spark training course, based on his bestselling *Taming Big Data with Apache Spark and Python* video, now available in a book. Understand and analyze large data sets using Spark on a single system or on a cluster. About This Book

Understand how Spark can be distributed across computing clusters
 Develop and run Spark jobs efficiently using Python
 A hands-on tutorial by Frank Kane with over 15 real-world examples teaching you Big Data processing with Spark
 Who This Book Is For If you are a data scientist or data analyst who wants to learn Big Data processing using Apache Spark and Python, this book is for you. If you have some programming experience in Python, and want to learn how to process large amounts of data using Apache Spark, Frank Kane's *Taming Big Data with Apache Spark and Python* will also help you.
 What You Will Learn Find out how you can identify Big Data problems as Spark problems
 Install and run Apache Spark on your computer or on a cluster
 Analyze large data sets across many CPUs using Spark's Resilient Distributed Datasets
 Implement machine learning on Spark using the MLib library
 Process continuous streams of data in real time using the Spark streaming module
 Perform complex network analysis using Spark's GraphX library
 Use Amazon's Elastic MapReduce service to run

your Spark jobs on a cluster
 In Detail Frank Kane's *Taming Big Data with Apache Spark and Python* is your companion to learning Apache Spark in a hands-on manner. Frank will start you off by teaching you how to set up Spark on a single system or on a cluster, and you'll soon move on to analyzing large data sets using Spark RDD, and developing and running effective Spark jobs quickly using Python. Apache Spark has emerged as the next big thing in the Big Data domain – quickly rising from an ascending technology to an established superstar in just a matter of years. Spark allows you to quickly extract actionable insights from large amounts of data, on a real-time basis, making it an essential tool in many modern businesses. Frank has packed this book with over 15 interactive, fun-filled examples relevant to the real world, and he will empower you to understand the Spark ecosystem and implement production-grade real-time Spark projects with ease. Style and approach
 Frank Kane's *Taming Big Data with Apache Spark and Python* is a hands-on tutorial with over 15 real-

world examples carefully explained by Frank in a step-by-step manner. The examples vary in complexity, and you can move through them at your own pace.

[Next-Generation Big Data](#)
 "O'Reilly Media, Inc."

This book is a step-by-step guide for learning how to use Spark for different types of big-data analytics projects, including batch, interactive, graph, and stream data analysis as well as machine learning. It covers Spark core and its add-on libraries, including Spark SQL, Spark Streaming, GraphX, MLib, and Spark ML. *Big Data Analytics with Spark* shows you how to use Spark and leverage its easy-to-use features to increase your productivity. You learn to perform fast data analysis using its in-memory caching and advanced execution engine, employ in-memory computing capabilities for building high-performance machine learning and low-latency interactive analytics applications, and much more. Moreover, the book shows you how to use Spark as a single integrated platform for a variety of data processing tasks, including ETL pipelines, BI, live data

stream processing, graph analytics, and machine learning. The book also includes a chapter on Scala, the hottest functional programming language, and the language that underlies Spark. You'll learn the basics of functional programming in Scala, so that you can write Spark applications in it. What's more, Big Data Analytics with Spark provides an introduction to other big data technologies that are commonly used along with Spark, such as HDFS, Avro, Parquet, Kafka, Cassandra, HBase, Mesos, and so on. It also provides an introduction to machine learning and graph concepts. So the book is self-sufficient; all the technologies that you need to know to use Spark are covered. The only thing that you are expected to have is some programming knowledge in any language.

Build highly effective analytics solutions to gain valuable insight into your big data

"O'Reilly Media, Inc."

Harness the power of Scala to program Spark and analyze tonnes of data in the blink of an eye! About This Book Learn Scala's sophisticated type system that combines Functional

Programming and object-oriented concepts Work on a wide array of applications, from simple batch jobs to stream processing and machine learning Explore the most common as well as some complex use-cases to perform large-scale data analysis with Spark Who This Book Is For Anyone who wishes to learn how to perform data analysis by harnessing the power of Spark will find this book extremely useful. No knowledge of Spark or Scala is assumed, although prior programming experience (especially with other JVM languages) will be useful to pick up concepts quicker. What You Will Learn Understand object-oriented & functional programming concepts of Scala In-depth understanding of Scala collection APIs Work with RDD and DataFrame to learn Spark's core abstractions Analysing structured and unstructured data using SparkSQL and GraphX Scalable and fault-tolerant streaming application development using Spark structured streaming Learn machine-learning best practices for classification, regression, dimensionality reduction, and recommendation

system to build predictive models with widely used algorithms in Spark MLlib & ML Build clustering models to cluster a vast amount of data Understand tuning, debugging, and monitoring Spark applications Deploy Spark applications on real clusters in Standalone, Mesos, and YARN In Detail Scala has been observing wide adoption over the past few years, especially in the field of data science and analytics. Spark, built on Scala, has gained a lot of recognition and is being used widely in productions. Thus, if you want to leverage the power of Scala and Spark to make sense of big data, this book is for you. The first part introduces you to Scala, helping you understand the object-oriented and functional programming concepts needed for Spark application development. It then moves on to Spark to cover the basic abstractions using RDD and DataFrame. This will help you develop scalable and fault-tolerant streaming applications by analyzing structured and unstructured data using SparkSQL, GraphX, and Spark structured streaming. Finally, the book moves on to some

advanced topics, such as monitoring, configuration, debugging, testing, and deployment. You will also learn how to develop Spark applications using SparkR and PySpark APIs, interactive data analytics using Zeppelin, and in-memory data processing with Alluxio. By the end of this book, you will have a thorough understanding of Spark, and you will be able to perform full-stack data analytics with a feel that no amount of data is too big. Style and approach Filled with practical examples and use cases, this book will not only help you get up and running with Spark, but will also take you farther down the road to becoming a data scientist.

Learning PySpark

Apress

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

A beginner's guide to harnessing the power and ease of PySpark 3

Packt Publishing Ltd
The Complete Guide to Data Science with Hadoop—For Technical Professionals, Businesspeople, and Students Demand is

soaring for professionals who can solve real data science problems with Hadoop and Spark. Practical Data Science with Hadoop® and Spark is your complete guide to doing just that. Drawing on immense experience with Hadoop and big data, three leading experts bring together everything you need: high-level concepts, deep-dive techniques, real-world use cases, practical applications, and hands-on tutorials. The authors introduce the essentials of data science and the modern Hadoop ecosystem, explaining how Hadoop and Spark have evolved into an effective platform for solving data science problems at scale. In addition to comprehensive application coverage, the authors also provide useful guidance on the important steps of data ingestion, data munging, and visualization. Once the groundwork is in place, the authors focus on specific applications, including machine learning, predictive modeling for sentiment analysis, clustering for document analysis, anomaly detection, and natural language processing (NLP). This

guide provides a strong technical foundation for those who want to do practical data science, and also presents business-driven guidance on how to apply Hadoop and Spark to optimize ROI of data science initiatives. Learn What data science is, how it has evolved, and how to plan a data science career How data volume, variety, and velocity shape data science use cases Hadoop and its ecosystem, including HDFS, MapReduce, YARN, and Spark Data importation with Hive and Spark Data quality, preprocessing, preparation, and modeling Visualization: surfacing insights from huge data sets Machine learning: classification, regression, clustering, and anomaly detection Algorithms and Hadoop tools for predictive modeling Cluster analysis and similarity functions Large-scale anomaly detection NLP: applying data science to human language FT Press Apache Spark is amazing when everything clicks. But if you haven't seen the performance improvements you expected, or still don'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, you'll also learn how to make it sing. With this book, you'll explore: How Spark SQL's new interfaces improve performance over SQL's RDD data structure The choice between data joins in Core Spark and Spark SQL Techniques for getting the most out of standard RDD transformations How to work around performance issues in Spark's key/value pair paradigm Writing high-performance Spark code without Scala or the JVM How to test for functionality and performance when applying suggested improvements Using Spark MLlib and Spark ML

machine learning libraries Spark's Streaming components and external community packages [Mastering Spark with R](#) Apress Summary Spark in Action teaches you the theory and skills you need to effectively handle batch and streaming data using Spark. Fully updated for Spark 2.0. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Big data systems distribute datasets across clusters of machines, making it a challenge to efficiently query, stream, and interpret them. Spark can help. It is a processing system designed specifically for distributed data. It provides easy-to-use interfaces, along with the performance you need for production-quality analytics and machine learning. Spark 2 also adds improved programming APIs, better performance, and countless other upgrades. About the Book Spark in Action teaches you the theory and skills you need to effectively handle batch and streaming data using Spark. You'll get comfortable with the Spark CLI as you work through a few

introductory examples. Then, you'll start programming Spark using its core APIs. Along the way, you'll work with structured data using Spark SQL, process near-real-time streaming data, apply machine learning algorithms, and munge graph data using Spark GraphX. For a zero-effort startup, you can download the preconfigured virtual machine ready for you to try the book's code.

What's Inside Updated for Spark 2.0 Real-life case studies Spark DevOps with Docker Examples in Scala, and online in Java and Python About the Reader Written for experienced programmers with some background in big data or machine learning. About the Authors Petar Zečević and Marko Bonaći are seasoned developers heavily involved in the Spark community. Table of Contents PART 1 - FIRST STEPS Introduction to Apache Spark Spark fundamentals Writing Spark applications The Spark API in depth PART 2 - MEET THE SPARK FAMILY Sparkling queries with Spark SQL Ingesting data with Spark Streaming Getting smart with MLlib ML: classification and clustering Connecting the dots with GraphX PART 3 -

SPARK OPS Running Spark Running on a Spark standalone cluster Running on YARN and Mesos PART 4 - BRINGING IT TOGETHER Case study: real-time dashboard Deep learning on Spark with H2O *Patterns for Learning from Data at Scale* "O'Reilly Media, Inc." Utilize R to uncover hidden patterns in your Big Data About This Book Perform computational analyses on Big Data to generate meaningful results Get a practical knowledge of R programming language while working on Big Data platforms like Hadoop, Spark, H2O and SQL/NoSQL databases, Explore fast, streaming, and scalable data analysis with the most cutting-edge technologies in the market Who This Book Is For This book is intended for Data Analysts, Scientists, Data Engineers, Statisticians, Researchers, who want to integrate R with their current or future Big Data workflows. It is assumed that readers have some experience in data analysis and understanding of data management and algorithmic processing of large quantities of data, however they may lack

specific skills related to R. What You Will Learn Learn about current state of Big Data processing using R programming language and its powerful statistical capabilities Deploy Big Data analytics platforms with selected Big Data tools supported by R in a cost-effective and time-saving manner Apply the R language to real-world Big Data problems on a multi-node Hadoop cluster, e.g. electricity consumption across various socio-demographic indicators and bike share scheme usage Explore the compatibility of R with Hadoop, Spark, SQL and NoSQL databases, and H2O platform In Detail Big Data analytics is the process of examining large and complex data sets that often exceed the computational capabilities. R is a leading programming language of data science, consisting of powerful functions to tackle all problems related to Big Data processing. The book will begin with a brief introduction to the Big Data world and its current industry standards. With introduction to the R language and presenting its development, structure, applications in real world, and its

shortcomings. Book will progress towards revision of major R functions for data management and transformations. Readers will be introduced to Cloud based Big Data solutions (e.g. Amazon EC2 instances and Amazon RDS, Microsoft Azure and its HDInsight clusters) and also provide guidance on R connectivity with relational and non-relational databases such as MongoDB and HBase etc. It will further expand to include Big Data tools such as Apache Hadoop ecosystem, HDFS and MapReduce frameworks. Also other R compatible tools such as Apache Spark, its machine learning library Spark MLlib, as well as H2O. Style and approach This book will serve as a practical guide to tackling Big Data problems using R programming language

and its statistical environment. Each section of the book will present you with concise and easy-to-follow steps on how to process, transform and analyse large data sets. Big Data Analytics with Hadoop 3 Packt Publishing Ltd
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

Related with Big Data Analytics With Spark Home Springer:

- Lindsay Lohan Dating History : [click here](#)