

# Practical Data Science With Hadoop And Spark

As recognized, adventure as with ease as experience approximately lesson, amusement, as with ease as promise can be gotten by just checking out a book **Practical Data Science With Hadoop And Spark** after that it is not directly done, you could receive even more with reference to this life, in this area the world.

We come up with the money for you this proper as with ease as easy pretension to acquire those all. We meet the expense of Practical Data Science With Hadoop And Spark and numerous books collections from fictions to scientific research in any way. along with them is this Practical Data Science With Hadoop And Spark that can be your partner.

*Next-Generation Big Data* - Butch Quinto  
2018-06-12

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** - Holden Karau 2015-01-28  
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

[Big Data Analytics with Hadoop 3](#) - Sridhar Alla  
2018-05-31

Explore big data concepts, platforms, analytics, and their applications using the power of Hadoop 3 Key Features Learn Hadoop 3 to build effective big data analytics solutions on-premise and on cloud Integrate Hadoop with other big data tools such as R, Python, Apache Spark, and Apache Flink Exploit big data using Hadoop 3 with real-world examples Book Description Apache Hadoop is the most popular platform for big data processing, and can be combined with a host of other big data tools to build powerful analytics solutions. Big Data Analytics with Hadoop 3 shows you how to do just that, by providing insights into the software as well as its benefits with the help of practical examples. Once you have taken a tour of Hadoop 3's latest features, you will get an overview of HDFS, MapReduce, and YARN, and how they enable faster, more efficient big data processing. You will then move on to learning how to integrate Hadoop with the open source tools, such as Python and R, to analyze and visualize data and perform statistical computing on big data. As you get acquainted with all this, you will explore how to use Hadoop 3 with Apache Spark and Apache Flink for real-time data analytics and stream processing. In addition to this, you will understand how to use Hadoop to build analytics solutions on the cloud and an end-to-end pipeline to perform big data analysis using practical use cases. By the end of this book, you will be well-versed with the analytical capabilities of the Hadoop ecosystem. You will be able to build powerful solutions to perform big data analytics and get insight

effortlessly. What you will learn Explore the new features of Hadoop 3 along with HDFS, YARN, and MapReduce Get well-versed with the analytical capabilities of Hadoop ecosystem using practical examples Integrate Hadoop with R and Python for more efficient big data processing Learn to use Hadoop with Apache Spark and Apache Flink for real-time data analytics Set up a Hadoop cluster on AWS cloud Perform big data analytics on AWS using Elastic Map Reduce Who this book is for Big Data Analytics with Hadoop 3 is for you if you are looking to build high-performance analytics solutions for your enterprise or business using Hadoop 3's powerful features, or you're new to big data analytics. A basic understanding of the Java programming language is required.

*Mastering Spark with R* - Javier Luraschi  
2019-10-07

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

**Applied Text Analysis with Python** - Benjamin Bengfort  
2018-06-11

From news and speeches to informal chatter on social media, natural language is one of the richest and most underutilized sources of data.

Not only does it come in a constant stream, always changing and adapting in context; it also contains information that is not conveyed by traditional data sources. The key to unlocking natural language is through the creative application of text analytics. This practical book presents a data scientist's approach to building language-aware products with applied machine learning. You'll learn robust, repeatable, and scalable techniques for text analysis with Python, including contextual and linguistic feature engineering, vectorization, classification, topic modeling, entity resolution, graph analysis, and visual steering. By the end of the book, you'll be equipped with practical methods to solve any number of complex real-world problems. Preprocess and vectorize text into high-dimensional feature representations Perform document classification and topic modeling Steer the model selection process with visual diagnostics Extract key phrases, named entities, and graph structures to reason about data in text Build a dialog framework to enable chatbots and language-driven interaction Use Spark to scale processing power and neural networks to scale model complexity

**Big Data Processing with Apache Spark** - Sрни Penchikala 2018-03-13

Apache Spark is a popular open-source big-data processing framework that's built around speed, ease of use, and unified distributed computing architecture. Not only it supports developing applications in different languages like Java, Scala, Python, and R, it's also hundred times faster in memory and ten times faster even when running on disk compared to traditional data processing frameworks. Whether you are currently working on a big data project or interested in learning more about topics like machine learning, streaming data processing, and graph data analytics, this book is for you. You can learn about Apache Spark and develop Spark programs for various use cases in big data analytics using the code examples provided. This book covers all the libraries in Spark ecosystem: Spark Core, Spark SQL, Spark Streaming, Spark ML, and Spark GraphX.

[Advanced Analytics with Spark](#) - Sandy Ryza 2015-04-14

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

**Practical Data Science Cookbook** -

Prabhanjan Tattar 2017-06-29

Over 85 recipes to help you complete real-world data science projects in R and Python About This Book Tackle every step in the data science pipeline and use it to acquire, clean, analyze, and visualize your data Get beyond the theory and implement real-world projects in data science using R and Python Easy-to-follow recipes will help you understand and implement the numerical computing concepts Who This Book Is For If you are an aspiring data scientist who wants to learn data science and numerical programming concepts through hands-on, real-world project examples, this is the book for you. Whether you are brand new to data science or you are a seasoned expert, you will benefit from learning about the structure of real-world data science projects and the programming examples in R and Python. What You Will Learn Learn and understand the installation procedure and environment required for R and Python on various platforms Prepare data for analysis by implement various data science concepts such as acquisition, cleaning and munging through R and Python Build a predictive model and an

exploratory model Analyze the results of your model and create reports on the acquired data Build various tree-based methods and Build random forest In Detail As increasing amounts of data are generated each year, the need to analyze and create value out of it is more important than ever. Companies that know what to do with their data and how to do it well will have a competitive advantage over companies that don't. Because of this, there will be an increasing demand for people that possess both the analytical and technical abilities to extract valuable insights from data and create valuable solutions that put those insights to use. Starting with the basics, this book covers how to set up your numerical programming environment, introduces you to the data science pipeline, and guides you through several data projects in a step-by-step format. By sequentially working through the steps in each chapter, you will quickly familiarize yourself with the process and learn how to apply it to a variety of situations with examples using the two most popular programming languages for data analysis—R and Python. Style and approach This step-by-step guide to data science is full of hands-on examples of real-world data science tasks. Each recipe focuses on a particular task involved in the data science pipeline, ranging from readying the dataset to analytics and visualization

[Spark for Data Science Cookbook - 2016](#)

### **Apache Spark for Data Science Cookbook -**

Padma Priya Chitturi 2016-12-22

Over insightful 90 recipes to get lightning-fast analytics with Apache Spark About This Book Use Apache Spark for data processing with these hands-on recipes Implement end-to-end, large-scale data analysis better than ever before Work with powerful libraries such as MLLib, SciPy, NumPy, and Pandas to gain insights from your data Who This Book Is For This book is for novice and intermediate level data science professionals and data analysts who want to solve data science problems with a distributed computing framework. Basic experience with data science implementation tasks is expected. Data science professionals looking to skill up and gain an edge in the field will find this book helpful. What You Will Learn Explore the topics of data mining, text mining, Natural Language Processing,

information retrieval, and machine learning. Solve real-world analytical problems with large data sets. Address data science challenges with analytical tools on a distributed system like Spark (apt for iterative algorithms), which offers in-memory processing and more flexibility for data analysis at scale. Get hands-on experience with algorithms like Classification, regression, and recommendation on real datasets using Spark MLLib package. Learn about numerical and scientific computing using NumPy and SciPy on Spark. Use Predictive Model Markup Language (PMML) in Spark for statistical data mining models. In Detail Spark has emerged as the most promising big data analytics engine for data science professionals. The true power and value of Apache Spark lies in its ability to execute data science tasks with speed and accuracy. Spark's selling point is that it combines ETL, batch analytics, real-time stream analysis, machine learning, graph processing, and visualizations. It lets you tackle the complexities that come with raw unstructured data sets with ease. This guide will get you comfortable and confident performing data science tasks with Spark. You will learn about implementations including distributed deep learning, numerical computing, and scalable machine learning. You will be shown effective solutions to problematic concepts in data science using Spark's data science libraries such as MLLib, Pandas, NumPy, SciPy, and more. These simple and efficient recipes will show you how to implement algorithms and optimize your work. Style and approach This book contains a comprehensive range of recipes designed to help you learn the fundamentals and tackle the difficulties of data science. This book outlines practical steps to produce powerful insights into Big Data through a recipe-based approach.

*Mastering Spark for Data Science - Andrew Morgan 2017-01-31*

Unlock the complexities of lightning fast data science About This Book\*Develop and apply advanced analytical techniques with Spark\*Learn how to tell a compelling story in data science using Spark's ecosystem\*Explore data at a scale and work with cutting edge data science methods Who This Book Is For This book is for those who have beginner-level familiarity with the Spark architecture and data science applications, who are looking for a challenge and

want to learn cutting edge techniques. This book assumes working knowledge of data science, common machine learning methods, and popular data science tools, and assumes you have previously run proof of concept studies and built prototypes. What You Will Learn\* Learn the design patterns that integrate Spark into with industrialized data science pipelines\* Understand how commercial data scientists design scalable code and reusable code for data science services\* Get a grasp of the new cutting edge data science methods so you can study trends and causality\* Find out how to use Spark as a universal ingestion engine tool and as a web scraper\* Practice the implementation of advanced topics in graph processing, such as community detection and contact chaining\* Get to know the best practices when performing Extended Exploratory Data Analysis, commonly used in commercial data science teams\* Grasp advanced Spark concepts, as well as solution design patterns and integration architectures\* Demonstrate powerful data science pipelines\* Get detailed guidance on how to run Spark in production

In Detail The purpose of data science is to transform the world using data, and this goal is mainly achieved through disrupting and changing real processes in real industries. To operate at this level, you need to be able to build data science solutions of substance; ones that solve real problems, and that can run reliably enough for people to trust and act on. Spark has emerged as the big data platform of choice for data scientists. This book deep dives into Spark to deliver production-grade data science solutions that are innovative, disruptive, and reliable enough to be trusted. We demonstrate the process through exploring the construction of a sophisticated global news analysis service that uses Spark to generate continuous geopolitical and current affairs insights. We use the core Spark APIs and take a deep-dive into advanced libraries including: Spark SQL, visual streaming, MLlib, and more. We introduce advanced techniques and methods to help you build data science solutions, and show you how to construct commercial grade data products. Using a sequence of tutorials that deliver a working news intelligence service, we explain advanced Spark architectures, unveil sophisticated data science methods, demonstrate how to work with

geographic data in Spark, and explain how to tune Spark algorithms so they scale linearly.

**Data Algorithms** - Mahmoud Parsian  
2015-07-13

If you are ready to dive into the MapReduce framework for processing large datasets, this practical book takes you step by step through the algorithms and tools you need to build distributed MapReduce applications with Apache Hadoop or Apache Spark. Each chapter provides a recipe for solving a massive computational problem, such as building a recommendation system. You'll learn how to implement the appropriate MapReduce solution with code that you can use in your projects. Dr. Mahmoud Parsian covers basic design patterns, optimization techniques, and data mining and machine learning solutions for problems in bioinformatics, genomics, statistics, and social network analysis. This book also includes an overview of MapReduce, Hadoop, and Spark. Topics include: Market basket analysis for a large set of transactions Data mining algorithms (K-means, KNN, and Naive Bayes) Using huge genomic data to sequence DNA and RNA Naive Bayes theorem and Markov chains for data and market prediction Recommendation algorithms and pairwise document similarity Linear regression, Cox regression, and Pearson correlation Allelic frequency and mining DNA Social network analysis (recommendation systems, counting triangles, sentiment analysis)

**Data Analytics with Spark Using Python** - Jeffrey Aven  
2018-06-18

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

*Practical Data Science with SAP* - Greg Foss  
2019-09-18

Learn how to fuse today's data science tools and techniques with your SAP enterprise resource planning (ERP) system. With this practical guide, SAP veterans Greg Foss and Paul Modderman demonstrate how to use several data analysis tools to solve interesting problems with your SAP data. Data engineers and scientists will explore ways to add SAP data to their analysis processes, while SAP business analysts will learn practical methods for answering questions about the business. By focusing on grounded explanations of both SAP processes and data science tools, this book gives data scientists and business analysts powerful methods for discovering deep data truths. You'll explore: Examples of how data analysis can help you solve several SAP challenges Natural language processing for unlocking the secrets in text Data science techniques for data clustering and segmentation Methods for detecting anomalies in your SAP data Data visualization techniques for making your data come to life

**Data Engineering with Apache Spark, Delta Lake, and Lakehouse** - Manoj Kukreja  
2021-10-22

Understand the complexities of modern-day data

engineering platforms and explore strategies to deal with them with the help of use case scenarios led by an industry expert in big data

**Key Features** Become well-versed with the core concepts of Apache Spark and Delta Lake for building data platforms Learn how to ingest, process, and analyze data that can be later used for training machine learning models Understand how to operationalize data models in production using curated data

**Book Description** In the world of ever-changing data and schemas, it is important to build data pipelines that can auto-adjust to changes. This book will help you build scalable data platforms that managers, data scientists, and data analysts can rely on. Starting with an introduction to data engineering, along with its key concepts and architectures, this book will show you how to use Microsoft Azure Cloud services effectively for data engineering. You'll cover data lake design patterns and the different stages through which the data needs to flow in a typical data lake. Once you've explored the main features of Delta Lake to build data lakes with fast performance and governance in mind, you'll advance to implementing the lambda architecture using Delta Lake. Packed with practical examples and code snippets, this book takes you through real-world examples based on production scenarios faced by the author in his 10 years of experience working with big data. Finally, you'll cover data lake deployment strategies that play an important role in provisioning the cloud resources and deploying the data pipelines in a repeatable and continuous way. By the end of this data engineering book, you'll know how to effectively deal with ever-changing data and create scalable data pipelines to streamline data science, ML, and artificial intelligence (AI) tasks. What you will learn

Discover the challenges you may face in the data engineering world Add ACID transactions to Apache Spark using Delta Lake Understand effective design strategies to build enterprise-grade data lakes Explore architectural and design patterns for building efficient data ingestion pipelines Orchestrate a data pipeline for preprocessing data using Apache Spark and Delta Lake APIs Automate deployment and monitoring of data pipelines in production Get to grips with securing, monitoring, and managing data pipelines models efficiently Who this book is

for This book is for aspiring data engineers and data analysts who are new to the world of data engineering and are looking for a practical guide to building scalable data platforms. If you already work with PySpark and want to use Delta Lake for data engineering, you'll find this book useful. Basic knowledge of Python, Spark, and SQL is expected.

**Spark for Data Science** - Srinivas Duvvuri  
2016-09-30

Analyze your data and delve deep into the world of machine learning with the latest Spark version, 2.0 About This Book Perform data analysis and build predictive models on huge datasets that leverage Apache Spark Learn to integrate data science algorithms and techniques with the fast and scalable computing features of Spark to address big data challenges Work through practical examples on real-world problems with sample code snippets Who This Book Is For This book is for anyone who wants to leverage Apache Spark for data science and machine learning. If you are a technologist who wants to expand your knowledge to perform data science operations in Spark, or a data scientist who wants to understand how algorithms are implemented in Spark, or a newbie with minimal development experience who wants to learn about Big Data Analytics, this book is for you! What You Will Learn Consolidate, clean, and transform your data acquired from various data sources Perform statistical analysis of data to find hidden insights Explore graphical techniques to see what your data looks like Use machine learning techniques to build predictive models Build scalable data products and solutions Start programming using the RDD, DataFrame and Dataset APIs Become an expert by improving your data analytical skills In Detail This is the era of Big Data. The words 'Big Data' implies big innovation and enables a competitive advantage for businesses. Apache Spark was designed to perform Big Data analytics at scale, and so Spark is equipped with the necessary algorithms and supports multiple programming languages. Whether you are a technologist, a data scientist, or a beginner to Big Data analytics, this book will provide you with all the skills necessary to perform statistical data analysis, data visualization, predictive modeling, and build scalable data products or solutions using Python, Scala, and R. With ample case

studies and real-world examples, Spark for Data Science will help you ensure the successful execution of your data science projects. Style and approach This book takes a step-by-step approach to statistical analysis and machine learning, and is explained in a conversational and easy-to-follow style. Each topic is explained sequentially with a focus on the fundamentals as well as the advanced concepts of algorithms and techniques. Real-world examples with sample code snippets are also included.

**Practical Real-time Data Processing and Analytics** - Shilpi Saxena 2017-09-28

A practical guide to help you tackle different real-time data processing and analytics problems using the best tools for each scenario About This Book Learn about the various challenges in real-time data processing and use the right tools to overcome them This book covers popular tools and frameworks such as Spark, Flink, and Apache Storm to solve all your distributed processing problems A practical guide filled with examples, tips, and tricks to help you perform efficient Big Data processing in real-time Who This Book Is For If you are a Java developer who would like to be equipped with all the tools required to devise an end-to-end practical solution on real-time data streaming, then this book is for you. Basic knowledge of real-time processing would be helpful, and knowing the fundamentals of Maven, Shell, and Eclipse would be great. What You Will Learn Get an introduction to the established real-time stack Understand the key integration of all the components Get a thorough understanding of the basic building blocks for real-time solution designing Garnish the search and visualization aspects for your real-time solution Get conceptually and practically acquainted with real-time analytics Be well equipped to apply the knowledge and create your own solutions In Detail With the rise of Big Data, there is an increasing need to process large amounts of data continuously, with a shorter turnaround time. Real-time data processing involves continuous input, processing and output of data, with the condition that the time required for processing is as short as possible. This book covers the majority of the existing and evolving open source technology stack for real-time processing and analytics. You will get to know about all the real-time solution aspects, from the source to the

presentation to persistence. Through this practical book, you'll be equipped with a clear understanding of how to solve challenges on your own. We'll cover topics such as how to set up components, basic executions, integrations, advanced use cases, alerts, and monitoring. You'll be exposed to the popular tools used in real-time processing today such as Apache Spark, Apache Flink, and Storm. Finally, you will put your knowledge to practical use by implementing all of the techniques in the form of a practical, real-world use case. By the end of this book, you will have a solid understanding of all the aspects of real-time data processing and analytics, and will know how to deploy the solutions in production environments in the best possible manner. **Style and Approach** In this practical guide to real-time analytics, each chapter begins with a basic high-level concept of the topic, followed by a practical, hands-on implementation of each concept, where you can see the working and execution of it. The book is written in a DIY style, with plenty of practical use cases, well-explained code examples, and relevant screenshots and diagrams.

**Practical Machine Learning** - Sunila Gollapudi  
2016-01-30

Tackle the real-world complexities of modern machine learning with innovative, cutting-edge, techniques **About This Book** Fully-coded working examples using a wide range of machine learning libraries and tools, including Python, R, Julia, and Spark **Comprehensive practical solutions** taking you into the future of machine learning **Go a step further and integrate your machine learning projects with Hadoop** **Who This Book Is For** This book has been created for data scientists who want to see machine learning in action and explore its real-world application. With guidance on everything from the fundamentals of machine learning and predictive analytics to the latest innovations set to lead the big data revolution into the future, this is an unmissable resource for anyone dedicated to tackling current big data challenges. Knowledge of programming (Python and R) and mathematics is advisable if you want to get started immediately. **What You Will Learn** Implement a wide range of algorithms and techniques for tackling complex data Get to grips with some of the most powerful languages in data science, including R, Python, and Julia

Harness the capabilities of Spark and Hadoop to manage and process data successfully Apply the appropriate machine learning technique to address real-world problems Get acquainted with Deep learning and find out how neural networks are being used at the cutting-edge of machine learning Explore the future of machine learning and dive deeper into polyglot persistence, semantic data, and more **In Detail** Finding meaning in increasingly larger and more complex datasets is a growing demand of the modern world. Machine learning and predictive analytics have become the most important approaches to uncover data gold mines. Machine learning uses complex algorithms to make improved predictions of outcomes based on historical patterns and the behaviour of data sets. Machine learning can deliver dynamic insights into trends, patterns, and relationships within data, immensely valuable to business growth and development. This book explores an extensive range of machine learning techniques uncovering hidden tricks and tips for several types of data using practical and real-world examples. While machine learning can be highly theoretical, this book offers a refreshing hands-on approach without losing sight of the underlying principles. Inside, a full exploration of the various algorithms gives you high-quality guidance so you can begin to see just how effective machine learning is at tackling contemporary challenges of big data. This is the only book you need to implement a whole suite of open source tools, frameworks, and languages in machine learning. We will cover the leading data science languages, Python and R, and the underrated but powerful Julia, as well as a range of other big data platforms including Spark, Hadoop, and Mahout. **Practical Machine Learning** is an essential resource for the modern data scientists who want to get to grips with its real-world application. With this book, you will not only learn the fundamentals of machine learning but dive deep into the complexities of real world data before moving on to using Hadoop and its wider ecosystem of tools to process and manage your structured and unstructured data. You will explore different machine learning techniques for both supervised and unsupervised learning; from decision trees to Naive Bayes classifiers and linear and clustering methods, you will learn strategies for a truly



advanced approach to the statistical analysis of data. The book also explores the cutting-edge advancements in machine learning, with worked examples and guidance on deep learning and reinforcement learning, providing you with practical demonstrations and samples that help take the theory-and mystery-out of even the most advanced machine learning methodologies. Style and approach A practical data science tutorial designed to give you an insight into the practical application of machine learning, this book takes you through complex concepts and tasks in an accessible way. Featuring information on a wide range of data science techniques, Practical Machine Learning is a comprehensive data science resource.

*Shu ju ke xue yu da shu ju ji shu dao lun* - □□□□  
2018

### **Big Data Analytics** - Venkat Ankam 2016-09-28

A handy reference guide for data analysts and data scientists to help to obtain value from big data analytics using Spark on Hadoop clusters

About This Book This book is based on the latest 2.0 version of Apache Spark and 2.7 version of Hadoop integrated with most commonly used tools. Learn all Spark stack components including latest topics such as DataFrames, DataSets, GraphFrames, Structured Streaming, DataFrame based ML Pipelines and SparkR. Integrations with frameworks such as HDFS, YARN and tools such as Jupyter, Zeppelin, NiFi, Mahout, HBase Spark Connector, GraphFrames, H2O and Hivemall.

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 and implement the tools and techniques of big data analytics using Spark on Hadoop clusters with wide variety of tools used with Spark and Hadoop Understand all the Hadoop and Spark ecosystem components Get to know all the Spark components: Spark Core, Spark SQL, DataFrames, DataSets, Conventional and Structured Streaming, MLLib, ML Pipelines and

Graphx See batch and real-time data analytics using Spark Core, Spark SQL, and Conventional and Structured Streaming Get to grips with data science and machine learning using MLLib, ML Pipelines, H2O, Hivemall, Graphx, SparkR and Hivemall. In Detail Big Data Analytics book aims at providing the fundamentals of Apache Spark and Hadoop. All Spark components – Spark Core, Spark SQL, DataFrames, Data sets, Conventional Streaming, Structured Streaming, MLLib, Graphx and Hadoop core components – HDFS, MapReduce and Yarn are explored in greater depth with implementation examples on Spark + Hadoop clusters. It is moving away from MapReduce to Spark. So, advantages of Spark over MapReduce are explained at great depth to reap benefits of in-memory speeds. DataFrames API, Data Sources API and new Data set API are explained for building Big Data analytical applications. Real-time data analytics using Spark Streaming with Apache Kafka and HBase is covered to help building streaming applications. New Structured streaming concept is explained with an IOT (Internet of Things) use case. Machine learning techniques are covered using MLLib, ML Pipelines and SparkR and Graph Analytics are covered with GraphX and GraphFrames components of Spark. Readers will also get an opportunity to get started with web based notebooks such as Jupyter, Apache Zeppelin and data flow tool Apache NiFi to analyze and visualize data. Style and approach This step-by-step pragmatic guide will make life easy no matter what your level of experience. You will deep dive into Apache Spark on Hadoop clusters through ample exciting real-life examples. Practical tutorial explains data science in simple terms to help programmers and data analysts get started with Data Science

Mathematical Problems in Data Science - Li M. Chen 2015-12-15

This book describes current problems in data science and Big Data. Key topics are data classification, Graph Cut, the Laplacian Matrix, Google Page Rank, efficient algorithms, hardness of problems, different types of big data, geometric data structures, topological data processing, and various learning methods. For unsolved problems such as incomplete data relation and reconstruction, the book includes possible solutions and both statistical and

computational methods for data analysis. Initial chapters focus on exploring the properties of incomplete data sets and partial-connectedness among data points or data sets. Discussions also cover the completion problem of Netflix matrix; machine learning method on massive data sets; image segmentation and video search. This book introduces software tools for data science and Big Data such MapReduce, Hadoop, and Spark. This book contains three parts. The first part explores the fundamental tools of data science. It includes basic graph theoretical methods, statistical and AI methods for massive data sets. In second part, chapters focus on the procedural treatment of data science problems including machine learning methods, mathematical image and video processing, topological data analysis, and statistical methods. The final section provides case studies on special topics in variational learning, manifold learning, business and financial data recovery, geometric search, and computing models. Mathematical Problems in Data Science is a valuable resource for researchers and professionals working in data science, information systems and networks. Advanced-level students studying computer science, electrical engineering and mathematics will also find the content helpful.

**Scala and Spark for Big Data Analytics** - Md. Rezaul Karim 2017-07-25

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 Analyzing 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.

**Practical Data Science with Hadoop and Spark** - Ofer Mendeleevitch 2016-12-08

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

**Big Data Analytics with Spark** - Mohammed Guller 2015-12-29

Big Data Analytics with Spark is a step-by-step guide for learning Spark, which is an open-source fast and general-purpose cluster computing framework for large-scale data analysis. You will learn 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. In addition, this book will help you become a much sought-after Spark expert. Spark is one of the hottest Big Data technologies. The amount of data generated

today by devices, applications and users is exploding. Therefore, there is a critical need for tools that can analyze large-scale data and unlock value from it. Spark is a powerful technology that meets that need. You can, for example, use Spark to perform low latency computations through the use of efficient caching and iterative algorithms; leverage the features of its shell for easy and interactive Data analysis; employ its fast batch processing and low latency features to process your real time data streams and so on. As a result, adoption of Spark is rapidly growing and is replacing Hadoop MapReduce as the technology of choice for big data analytics. This book provides an introduction to Spark and related big-data technologies. It covers Spark core and its add-on libraries, including Spark SQL, Spark Streaming, GraphX, and MLlib. Big Data Analytics with Spark is therefore written for busy professionals who prefer learning a new technology from a consolidated source instead of spending countless hours on the Internet trying to pick bits and pieces from different sources. The book also provides a chapter on Scala, the hottest functional programming language, and the program 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, like Hive, Avro, Kafka and so on. 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 know is programming in any language. There is a critical shortage of people with big data expertise, so companies are willing to pay top dollar for people with skills in areas like Spark and Scala. So reading this book and absorbing its principles will provide a boost—possibly a big boost—to your career. [Advanced Analytics with Spark](#) - Sandy Ryza 2017-06-12

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. You'll start with an introduction to Spark and its ecosystem, and then dive into patterns that apply common techniques—including classification, clustering, collaborative filtering, and anomaly detection—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 the book's patterns useful for working on your own data applications. With this book, you will:

- Familiarize yourself with the Spark programming model
- Become comfortable within the Spark ecosystem
- Learn general approaches in data science
- Examine complete implementations that analyze large public data sets
- Discover which machine learning tools make sense for particular problems
- Acquire code that can be adapted to many uses

*Big Data Analysis with Python* - Ivan Marin  
2019-04-10

Get to grips with processing large volumes of data and presenting it as engaging, interactive insights using Spark and Python. Key Features

- Get a hands-on, fast-paced introduction to the Python data science stack
- Explore ways to create useful metrics and statistics from large datasets
- Create detailed analysis reports with real-world data

**Book Description** Processing big data in real time is challenging due to scalability, information inconsistency, and fault tolerance. *Big Data Analysis with Python* teaches you how to use tools that can control this data avalanche for you. With this book, you'll learn practical techniques to aggregate data into useful dimensions for posterior analysis, extract statistical measurements, and transform datasets into features for other systems. The book begins with an introduction to data manipulation in Python using pandas. You'll then get familiar with statistical analysis and plotting techniques. With multiple hands-on activities in store, you'll be able to analyze data that is distributed on several computers by using Dask. As you progress, you'll study how to aggregate data for plots when the entire data cannot be accommodated in memory. You'll also explore Hadoop (HDFS and YARN), which will help you tackle larger datasets. The book also covers Spark and explains how it interacts with other tools. By the end of this

book, you'll be able to bootstrap your own Python environment, process large files, and manipulate data to generate statistics, metrics, and graphs. What you will learn

- Use Python to read and transform data into different formats
- Generate basic statistics and metrics using data on disk
- Work with computing tasks distributed over a cluster
- Convert data from various sources into storage or querying formats
- Prepare data for statistical analysis, visualization, and machine learning
- Present data in the form of effective visuals

Who this book is for *Big Data Analysis with Python* is designed for Python developers, data analysts, and data scientists who want to get hands-on with methods to control data and transform it into impactful insights. Basic knowledge of statistical measurements and relational databases will help you to understand various concepts explained in this book.

**Practical Big Data Analytics** - Nataraj Dasgupta  
2018-01-15

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.

**Learning PySpark** - Tomasz Drabas 2017-02-27 Build data-intensive applications locally and deploy at scale using the combined powers of Python and Spark 2.0 About This Book Learn why and how you can efficiently use Python to process data and build machine learning models in Apache Spark 2.0 Develop and deploy efficient, scalable real-time Spark solutions Take your understanding of using Spark with Python to the next level with this jump start guide Who This Book Is For If you are a Python developer who wants to learn about the Apache Spark 2.0 ecosystem, this book is for you. A firm understanding of Python is expected to get the best out of the book. Familiarity with Spark would be useful, but is not mandatory. What You Will Learn Learn about Apache Spark and the Spark

2.0 architecture Build and interact with Spark DataFrames using Spark SQL Learn how to solve graph and deep learning problems using GraphFrames and TensorFrames respectively Read, transform, and understand data and use it to train machine learning models Build machine learning models with MLlib and ML Learn how to submit your applications programmatically using spark-submit Deploy locally built applications to a cluster In Detail Apache Spark is an open source framework for efficient cluster computing with a strong interface for data parallelism and fault tolerance. This book will show you how to leverage the power of Python and put it to use in the Spark ecosystem. You will start by getting a firm understanding of the Spark 2.0 architecture and how to set up a Python environment for Spark. You will get familiar with the modules available in PySpark. You will learn how to abstract data with RDDs and DataFrames and understand the streaming capabilities of PySpark. Also, you will get a thorough overview of machine learning capabilities of PySpark using ML and MLlib, graph processing using GraphFrames, and polyglot persistence using Blaze. Finally, you will learn how to deploy your applications to the cloud using the spark-submit command. By the end of this book, you will have established a firm understanding of the Spark Python API and how it can be used to build data-intensive applications. Style and approach This book takes a very comprehensive, step-by-step approach so you understand how the Spark ecosystem can be used with Python to develop efficient, scalable solutions. Every chapter is standalone and written in a very easy-to-understand manner, with a focus on both the hows and the whys of each concept.

DIY Data Science - Jonathan Dinu 2017-05-23 Data science underlies Amazon's product recommender, LinkedIn's People You Know feature, Pandora's personalized radio stations, Stripe's fraud detectors, and the incredible insights arising from the world's increasingly ubiquitous sensors. In the future, the world's most interesting and impactful problems will be solved with data science. But right now, there's a shortage of data scientists in every industry, traditional schools can't teach students fast enough, and much of the knowledge data scientists need remains trapped in large tech

companies. This comprehensive, practical tutorial is the solution. Drawing on his experience building Zipfian Academy's immersive 12-week data science training program, Jonathan Dinu brings together all you need to teach yourself data science, and successfully enter the profession. First, Dinu helps you internalize the data science "mindset": that virtually anything can be quantified, and once you have data, you can harvest amazing insights through statistical analysis and machine learning. He illuminates data science as it really is: a holistic, interdisciplinary process that encompasses the collection, processing, and communication of data: all that data scientists do, say, and believe. With this foundation in place, he teaches core data science skills through hands-on Python and SQL-based exercises integrated with a full book-length case study. Step by step, you'll learn how to leverage algorithmic thinking and the power of code, gain intuition about the power and limitations of current machine learning methods, and effectively apply them to real business problems. You'll walk through: Building basic and advanced models Performing exploratory data analysis Using data analysis to acquire and retain users or customers Making predictions with regression Using machine learning techniques Working with unsupervised learning and NLP Communicating with data Performing social network analyses Working with data at scale Getting started with Hadoop, Spark and other advanced tools Recognizing where common approaches break down, and how to overcome real world constraints Taking your next steps in your study and career Well-crafted appendices provide reference material on everything from the basics of Python and SQL to the essentials of probability, statistics, and linear algebra -- even preparing for your data science job interview!

**Big-Data Analytics for Cloud, IoT and Cognitive Computing** - Kai Hwang 2017-08-14

The definitive guide to successfully integrating social, mobile, Big-Data analytics, cloud and IoT principles and technologies The main goal of this book is to spur the development of effective big-data computing operations on smart clouds that are fully supported by IoT sensing, machine learning and analytics systems. To that end, the authors draw upon their original research and proven track record in the field to describe a

practical approach integrating big-data theories, cloud design principles, Internet of Things (IoT) sensing, machine learning, data analytics and Hadoop and Spark programming. Part 1 focuses on data science, the roles of clouds and IoT devices and frameworks for big-data computing. Big data analytics and cognitive machine learning, as well as cloud architecture, IoT and cognitive systems are explored, and mobile cloud-IoT-interaction frameworks are illustrated with concrete system design examples. Part 2 is devoted to the principles of and algorithms for machine learning, data analytics and deep learning in big data applications. Part 3 concentrates on cloud programming software libraries from MapReduce to Hadoop, Spark and TensorFlow and describes business, educational, healthcare and social media applications for those tools. The first book describing a practical approach to integrating social, mobile, analytics, cloud and IoT (SMACT) principles and technologies Covers theory and computing techniques and technologies, making it suitable for use in both computer science and electrical engineering programs Offers an extremely well-informed vision of future intelligent and cognitive computing environments integrating SMACT technologies Fully illustrated throughout with examples, figures and approximately 150 problems to support and reinforce learning Features a companion website with an instructor manual and PowerPoint slides [www.wiley.com/go/hwangIoT](http://www.wiley.com/go/hwangIoT) Big-Data Analytics for Cloud, IoT and Cognitive Computing satisfies the demand among university faculty and students for cutting-edge information on emerging intelligent and cognitive computing systems and technologies. Professionals working in data science, cloud computing and IoT applications will also find this book to be an extremely useful working resource.

**Data Analytics with Hadoop** - Benjamin Bengfort 2016-06-01

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

**Practical Data Science with Python** - Nathan George 2021-09-30

Learn to effectively manage data and execute data science projects from start to finish using Python Key Features Understand and utilize data science tools in Python, such as specialized machine learning algorithms and statistical modeling Build a strong data science foundation with the best data science tools available in Python Add value to yourself, your organization, and society by extracting actionable insights from raw data Book Description Practical Data Science with Python teaches you core data science concepts, with real-world and realistic examples, and strengthens your grip on the basic as well as advanced principles of data preparation and storage, statistics, probability theory, machine learning, and Python programming, helping you build a solid foundation to gain proficiency in data science. The book starts with an overview of basic Python skills and then introduces foundational data science techniques, followed by a thorough explanation of the Python code needed to execute the techniques. You'll understand the code by working through the examples. The code has been broken down into small chunks (a few lines or a function at a time) to enable thorough discussion. As you progress, you will learn how to

perform data analysis while exploring the functionalities of key data science Python packages, including pandas, SciPy, and scikit-learn. Finally, the book covers ethics and privacy concerns in data science and suggests resources for improving data science skills, as well as ways to stay up to date on new data science developments. By the end of the book, you should be able to comfortably use Python for basic data science projects and should have the skills to execute the data science process on any data source. What you will learn Use Python data science packages effectively Clean and prepare data for data science work, including feature engineering and feature selection Data modeling, including classic statistical models (such as t-tests), and essential machine learning algorithms, such as random forests and boosted models Evaluate model performance Compare and understand different machine learning methods Interact with Excel spreadsheets through Python Create automated data science reports through Python Get to grips with text analytics techniques Who this book is for The book is intended for beginners, including students starting or about to start a data science, analytics, or related program (e.g. Bachelor's, Master's, bootcamp, online courses), recent college graduates who want to learn new skills to set them apart in the job market, professionals who want to learn hands-on data science techniques in Python, and those who want to shift their career to data science. The book requires basic familiarity with Python. A "getting started with Python" section has been included to get complete novices up to speed.

*Handbook of IoT and Big Data* - Vijender Kumar Solanki 2019-02-21

This multi-contributed handbook focuses on the latest workings of IoT (internet of Things) and Big Data. As the resources are limited, it's the endeavor of the authors to support and bring the information into one resource. The book is divided into 4 sections that covers IoT and technologies, the future of Big Data, algorithms, and case studies showing IoT and Big Data in various fields such as health care, manufacturing and automation. Features Focuses on the latest workings of IoT and Big Data Discusses the emerging role of technologies and the fast-growing market of Big Data Covers the

movement toward automation with hardware, software, and sensors, and trying to save on energy resources Offers the latest technology on IoT Presents the future horizons on Big Data  
*Mastering Apache Spark 2.x* - Romeo Kienzler  
2017-07-26

Advanced analytics on your Big Data with latest Apache Spark 2.x About This Book An advanced guide with a combination of instructions and practical examples to extend the most up-to date Spark functionalities. Extend your data processing capabilities to process huge chunk of data in minimum time using advanced concepts in Spark. Master the art of real-time processing with the help of Apache Spark 2.x Who This Book Is For If you are a developer with some experience with Spark and want to strengthen your knowledge of how to get around in the world of Spark, then this book is ideal for you. Basic knowledge of Linux, Hadoop and Spark is assumed. Reasonable knowledge of Scala is expected. What You Will Learn Examine Advanced Machine Learning and DeepLearning with MLlib, SparkML, SystemML, H2O and DeepLearning4J Study highly optimised unified batch and real-time data processing using SparkSQL and Structured Streaming Evaluate large-scale Graph Processing and Analysis using GraphX and GraphFrames Apply Apache Spark in Elastic deployments using Jupyter and Zeppelin Notebooks, Docker, Kubernetes and the IBM Cloud Understand internal details of cost based optimizers used in Catalyst, SystemML and GraphFrames Learn how specific parameter settings affect overall performance of an Apache Spark cluster Leverage Scala, R and python for your data science projects In Detail Apache Spark is an in-memory cluster-based parallel processing system that provides a wide range of functionalities such as graph processing, machine learning, stream processing, and SQL. This book aims to take your knowledge of Spark to the next level by teaching you how to expand Spark's functionality and implement your data flows and machine/deep learning programs on top of the platform. The book commences with an overview of the Spark ecosystem. It will introduce you to Project Tungsten and Catalyst, two of the major advancements of Apache Spark 2.x. You will understand how memory management and binary processing, cache-

aware computation, and code generation are used to speed things up dramatically. The book extends to show how to incorporate H2O, SystemML, and DeepLearning4j for machine learning, and Jupyter Notebooks and Kubernetes/Docker for cloud-based Spark. During the course of the book, you will learn about the latest enhancements to Apache Spark 2.x, such as interactive querying of live data and unifying DataFrames and Datasets. You will also learn about the updates on the APIs and how DataFrames and Datasets affect SQL, machine learning, graph processing, and streaming. You will learn to use Spark as a big data operating system, understand how to implement advanced analytics on the new APIs, and explore how easy it is to use Spark in day-to-day tasks. Style and approach This book is an extensive guide to Apache Spark modules and tools and shows how Spark's functionality can be extended for real-time processing and storage with worked examples.

**Learning Pyspark** - Tomasz Drabas 2017-04-28  
Build real-time data intensive applications using the combined power of Python and Spark 2.0 About This Book\* Learn why and how you can efficiently use Python to implement various functionalities in Spark 2.0\* Develop efficient, scalable real-time Spark solutions and deploy them\* A comprehensive guide to take your understanding of implementing Spark with Python to the next level Who This Book Is For If you are a Python developer who wants to learn about the Spark 2.0 ecosystem and how its functionalities can be implemented in Python, this book is for you. A firm understanding of Python is expected to get the best out of the book. Familiarity with Spark would be useful, but is not mandatory. What you will learn\* Install, configure, and interact with Spark on a single machine\* Build and interact with Spark DataFrames and Datasets using Spark SQL abstraction\* Abstract various data sources with Blaze\* Read, transform, and understand data and use it to train machine learning models\* Familiarize yourself with the modeling pipeline capabilities of the machine learning module\* Build machine learning models with MLlib\* Package your application dependencies with spark-submit\* Deploy locally built applications to cluster In Detail Apache Spark is an open source



framework for efficient cluster computing with a strong interface for data parallelism and fault tolerance. This book will show you how you can leverage the power of Python and put it to use in the Spark ecosystem. You will start by getting a firm understanding of the Spark 2.0 architecture and how to set up a Python environment for Spark. Next you will get familiar with the PySpark packages and see how to get the data ready for processing. You will find out how to use the PySpark classes for RDD abstraction and Spark SQL abstraction, and understand streaming capabilities of Spark, machine learning using MLlib, polyglot persistence using Blaze, and graph processing using GraphX. Finally, you will see how you can configure Spark to deploy your applications on the cloud. By the end of this book, you will have established a firm understanding of the Spark Python API and how it can be used to build data-intensive applications.

Big Data Analytics with R - Simon Walkowiak  
2016-07-29

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.

Machine Learning with Spark - Rajdeep Dua  
2017-04-28

Create scalable machine learning applications to power a modern data-driven business using Spark 2.x About This Book Get to the grips with the latest version of Apache Spark Utilize Spark's machine learning library to implement predictive analytics Leverage Spark's powerful tools to load, analyze, clean, and transform your data Who This Book Is For If you have a basic knowledge of machine learning and want to implement various machine-learning concepts in the context of Spark ML, this book is for you. You should be well versed with the Scala and Python languages. What You Will Learn Get hands-on with the latest version of Spark ML Create your first Spark program with Scala and Python Set up and configure a development environment for Spark on your own computer, as well as on Amazon EC2 Access public machine learning datasets and

use Spark to load, process, clean, and transform data Use Spark's machine learning library to implement programs by utilizing well-known machine learning models Deal with large-scale text data, including feature extraction and using text data as input to your machine learning models Write Spark functions to evaluate the performance of your machine learning models In Detail This book will teach you about popular machine learning algorithms and their implementation. You will learn how various machine learning concepts are implemented in the context of Spark ML. You will start by installing Spark in a single and multinode cluster. Next you'll see how to execute Scala and Python based programs for Spark ML. Then we will take a few datasets and go deeper into clustering, classification, and regression. Toward the end, we will also cover text processing using Spark ML. Once you have learned the concepts, they can be applied to implement algorithms in either green-field implementations or to migrate existing systems to this new platform. You can migrate from Mahout or Scikit to use Spark ML. By the end of this book, you will acquire the skills to leverage Spark's features to create your own scalable machine learning applications and power a modern data-driven business. Style and approach This practical tutorial with real-world use cases enables you to develop your own machine learning systems with Spark. The examples will help you combine various techniques and models into an intelligent machine learning system.

### **Machine Learning with Spark - Second Edition** - Rajdeep Dua 2016-10-31

Develop intelligent machine learning systems with Spark About This Book \*Get the grips with the latest version of Apache Spark \*Utilize Spark's machine learning library to implement predictive analytics \*Leverage Spark's powerful tools to load, analyze, clean, and transform your data Who This Book Is For If you have a basic knowledge of machine learning and want to implement various machine-learning concepts in the context of Spark ML, this book is for you. You should be well versed with the Scala and Python languages. What You Will Learn \*Get hands-on with the latest version of Spark ML \*Create your first Spark program with Scala and Python \*Set up and configure a development environment for

Spark on your own computer, as well as on Amazon EC2 \*Access public machine learning datasets and use Spark to load, process, clean, and transform data \*Use Spark's machine learning library to implement programs by utilizing well-known machine learning models \*Deal with large-scale text data, including feature extraction and using text data as input to your machine learning models \*Write Spark functions to evaluate the performance of your machine learning models In Detail Spark ML is the machine learning module of Spark. It uses in-memory RDDs to process machine learning models faster for clustering, classification, and regression. This book will teach you about popular machine learning algorithms and their implementation. You will learn how various machine learning concepts are implemented in the context of Spark ML. You will start by installing Spark in a single and multinode cluster. Next you'll see how to execute Scala and Python based programs for Spark ML. Then we will take a few datasets and go deeper into clustering, classification, and regression. Toward the end, we will also cover text processing using Spark ML. Once you have learned the concepts, they can be applied to implement algorithms in either green-field implementations or to migrate existing systems to this new platform. You can migrate from Mahout or Scikit to use Spark ML.

### **Apache Spark Machine Learning Blueprints** - Alex Liu 2016-05-30

Develop a range of cutting-edge machine learning projects with Apache Spark using this actionable guide About This Book Customize Apache Spark and R to fit your analytical needs in customer research, fraud detection, risk analytics, and recommendation engine development Develop a set of practical Machine Learning applications that can be implemented in real-life projects A comprehensive, project-based guide to improve and refine your predictive models for practical implementation Who This Book Is For If you are a data scientist, a data analyst, or an R and SPSS user with a good understanding of machine learning concepts, algorithms, and techniques, then this is the book for you. Some basic understanding of Spark and its core elements and application is required. What You Will Learn Set up Apache Spark for machine learning and discover its impressive

processing power Combine Spark and R to unlock detailed business insights essential for decision making Build machine learning systems with Spark that can detect fraud and analyze financial risks Build predictive models focusing on customer scoring and service ranking Build a recommendation systems using SPSS on Apache Spark Tackle parallel computing and find out how it can support your machine learning projects Turn open data and communication data into actionable insights by making use of various forms of machine learning In Detail There's a reason why Apache Spark has become one of the most popular tools in Machine Learning - its ability to handle huge datasets at an impressive speed means you can be much more responsive to the data at your disposal. This book shows you Spark at its very best, demonstrating how to connect it with R and unlock maximum value not only from the tool but also from your data. Packed with a range of project "blueprints" that demonstrate some of the most interesting challenges that Spark can help you tackle, you'll find out how to use Spark notebooks and access, clean, and join different datasets before putting your knowledge into practice with some real-world projects, in which you will see how Spark Machine Learning can help you with everything from fraud detection to analyzing customer

attrition. You'll also find out how to build a recommendation engine using Spark's parallel computing powers. Style and approach This book offers a step-by-step approach to setting up Apache Spark, and use other analytical tools with it to process Big Data and build machine learning projects. The initial chapters focus more on the theory aspect of machine learning with Spark, while each of the later chapters focuses on building standalone projects using Spark.

### **Big Data Analytics - a Practical Approach -**

Venkatesh Gauri Shankar 2020-11-06

The book is organized in such a way that it covers the practical approaches and syllabus of Big Data Analytics prescribed by different universities and colleges for B.Tech./BE (Data Science), M.Tech./ME (Data Science), MSc. (Big Data Analytics), B.Tech (CSE), B.Tech (IT) and all other branches related to Big Data, ML and AI. This book offers a deep understanding of the essential practical concept of Big Data Analytics, including Hadoop, Spark and Tableau. This book is primarily designed to introduce some of the fundamental concepts with practical approaches of Big Data Analytics for undergraduate and postgraduate students in Data Science, CSE, and IT. The concepts in each chapter are explained coherently and unequivocally. Hence it enables the students to gain knowledge in the practical field of Big Data concepts with ease.