

Advanced Analytics With Spark Patterns For Learning From Data At Scale

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

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.

Combine advanced analytics including Machine Learning, Deep Learning Neural Networks and Natural Language Processing with modern scalable technologies including Apache Spark to derive actionable insights from Big Data in real-time Key Features Make a hands-on start in the fields of Big Data, Distributed Technologies and Machine Learning Learn how to design, develop and interpret the results of common Machine Learning algorithms Uncover hidden patterns in your data in order to derive real actionable insights and business value Book Description Every person and every organization in the world manages data, whether they realize it or not. Data is used to describe the world around us and can be used for almost any purpose, from analyzing consumer habits to fighting disease and serious organized crime. Ultimately, we manage data in order to derive value from it, and many organizations around the world have traditionally invested in technology to help process their data faster and more efficiently. But we now live in an interconnected world driven by mass data creation and consumption where data is no longer rows and columns restricted to a spreadsheet, but an organic and evolving asset in its own right. With this realization comes major challenges for organizations: how do we manage the sheer size of data being created every second (think not only spreadsheets and databases, but also social media posts, images, videos, music, blogs and so on)? And once we can manage all of this data, how do we derive real value from it? The focus of Machine Learning with Apache Spark is to help us answer these questions in a hands-on manner. We introduce the latest scalable technologies to help us manage and process big data. We then introduce advanced analytical algorithms applied to real-world use cases in order to uncover patterns, derive actionable insights, and learn from this big data. What you will learn Understand how Spark fits in the context of the big data ecosystem Understand how to deploy and configure a local development environment using Apache Spark Understand how to design supervised and unsupervised learning models Build models to perform NLP, deep learning, and cognitive services using Spark ML libraries Design real-time machine learning pipelines in Apache Spark Become familiar with advanced techniques for processing a large volume of data by applying machine learning algorithms Who this book is for This book is

aimed at Business Analysts, Data Analysts and Data Scientists who wish to make a hands-on start in order to take advantage of modern Big Data technologies combined with Advanced Analytics.

Applications of Big Data in Healthcare: Theory and Practice begins with the basics of Big Data analysis and introduces the tools, processes and procedures associated with Big Data analytics. The book unites healthcare with Big Data analysis and uses the advantages of the latter to solve the problems faced by the former. The authors present the challenges faced by the healthcare industry, including capturing, storing, searching, sharing and analyzing data. This book illustrates the challenges in the applications of Big Data and suggests ways to overcome them, with a primary emphasis on data repositories, challenges, and concepts for data scientists, engineers and clinicians. The applications of Big Data have grown tremendously within the past few years and its growth can not only be attributed to its competence to handle large data streams but also to its abilities to find insights from complex, noisy, heterogeneous, longitudinal and voluminous data. The main objectives of Big Data in the healthcare sector is to come up with ways to provide personalized healthcare to patients by taking into account the enormous amounts of already existing data. Provides case studies that illustrate the business processes underlying the use of big data and deep learning health analytics to improve health care delivery Supplies readers with a foundation for further specialized study in clinical analysis and data management Includes links to websites, videos, articles and other online content to expand and support the primary learning objectives for each major section of the book

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

Gain a broad foundation of advanced data analytics concepts and discover the recent revolution in databases such as Neo4j, Elasticsearch, and MongoDB. This book discusses how to implement ETL techniques including topical crawling, which is applied in domains such as high-frequency algorithmic trading and goal-oriented dialog systems. You'll also see examples of machine learning concepts such as semi-supervised learning, deep learning, and NLP. Advanced Data Analytics Using Python also covers important traditional data analysis techniques such as time series and principal component analysis. After reading this book you will have experience of every technical aspect of an analytics project. You'll get to know the concepts using Python code, giving you samples to use in your own projects. What You Will Learn Work with data analysis techniques such as classification, clustering, regression, and forecasting Handle structured and unstructured data, ETL techniques, and different kinds of databases such as Neo4j, Elasticsearch, MongoDB, and MySQL Examine the different big data frameworks, including Hadoop and Spark Discover advanced machine learning concepts such as semi-supervised learning, deep learning, and NLP Who This Book Is For Data scientists and software developers interested in the field of data analytics.

Leverage the power of Scala with different tools to build scalable, robust data science applications About This Book A complete guide for scalable data science solutions, from data ingestion to data visualization Deploy horizontally scalable data processing pipelines and take advantage of web frameworks to build engaging visualizations Build functional, type-safe routines to interact with relational and NoSQL databases with the help of tutorials and examples provided Who This Book Is For If you are a Scala developer or data scientist, or if you want to enter the field of data science, then this book will give you all the tools you need to implement data science solutions. What You Will Learn Transform and filter tabular data to extract features for machine learning Implement your own algorithms or take advantage of MLLib's extensive suite of models to build distributed machine learning pipelines Read, transform, and write data to both SQL and NoSQL databases in a functional manner Write robust routines to query web APIs Read data from web APIs such as the GitHub or Twitter API Use Scala to interact with MongoDB, which offers high performance and helps to store large data sets with uncertain query requirements Create Scala web applications that couple with JavaScript libraries such as D3 to create compelling interactive visualizations Deploy scalable parallel applications using Apache Spark, loading data from HDFS or Hive In Detail Scala is a multi-paradigm programming language (it supports both object-oriented and functional programming) and scripting language used to build applications for the JVM. Languages such as R, Python, Java, and so on are mostly used for data science. It is particularly good at analyzing large sets of data without any significant impact on performance and thus Scala is being adopted by many developers and data scientists. Data scientists might be aware that building applications that are truly scalable is hard. Scala, with its powerful functional libraries for interacting with databases and building scalable frameworks will give you the tools to construct robust data pipelines. This book will introduce you to the libraries for ingesting, storing, manipulating, processing, and visualizing data in Scala. Packed with real-world examples and interesting data sets, this book will teach you to ingest data from flat files and web APIs and store it in a SQL or NoSQL database. It will show you how to design scalable architectures to process and modelling your data, starting from simple concurrency constructs such as parallel collections and futures, through to actor systems and Apache Spark. As well as Scala's emphasis on functional structures and immutability, you will learn how to use the right parallel construct for the job at hand, minimizing development time without compromising scalability. Finally, you will learn how to build beautiful interactive visualizations using web frameworks. This book gives tutorials on some of the most common Scala libraries for data science, allowing you to quickly get up to speed with building data science and data engineering solutions. Style and approach A tutorial with complete examples, this book will give you the tools to start building useful data engineering and data science solutions straightaway 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

Learn how to use, deploy, and maintain Apache Spark with this comprehensive guide, written by the creators of the open-source cluster-computing framework. With an emphasis on improvements and new features in Spark 2.0, authors Bill Chambers and Matei Zaharia break down Spark topics into distinct sections, each with unique goals. You'll explore the basic operations and common functions of Spark's structured APIs, as well as Structured Streaming, a new high-level API for building end-to-end streaming applications. Developers and system administrators will learn the fundamentals of monitoring, tuning, and debugging Spark, and explore machine learning techniques and scenarios for employing MLlib, Spark's scalable machine-learning library. Get a gentle overview of big data and Spark Learn about DataFrames, SQL, and Datasets—Spark's core APIs—through worked examples Dive into Spark's low-level APIs, RDDs, and execution of SQL and DataFrames Understand how Spark runs on a cluster Debug, monitor, and tune Spark clusters and applications Learn the power of Structured Streaming, Spark's stream-processing engine Learn how you can apply MLlib to a variety of problems, including classification or recommendation

Advanced Analytics with Spark Patterns for Learning from Data at Scale"O'Reilly Media, Inc."

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

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.

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

The amount of data being generated today is staggering--and growing. Apache Spark has emerged as the de facto tool to analyze big data and is now a critical part of the data science toolbox. Updated for Spark 3.0, this practical guide brings together Spark, statistical methods, and real-world datasets to teach you how to approach analytics problems using PySpark, Spark's Python API, and other best practices in Spark programming. Data scientists Akash Tandon, Sandy Ryza, Uri Laserson, Sean Owen, and Josh Wills offer an introduction to the Spark ecosystem, then dive into patterns that apply common techniques--including classification, clustering, collaborative filtering, and anomaly detection--to fields such as genomics, security, and finance. This updated edition also covers NLP and image processing. If you have a basic understanding of machine learning and statistics and you program in Python, this book will get you started with large-scale data analysis. Familiarize yourself with Spark's programming model and ecosystem Learn general approaches in data science Examine complete implementations that analyze large public datasets Discover which machine learning tools make sense for particular problems Explore code that can be adapted to many uses

Master the techniques and sophisticated analytics used to construct Spark-based solutions that scale to deliver production-grade data science products About This Book Develop and apply advanced analytical techniques with Spark Learn how to tell a compelling story with data science using Spark's ecosystem Explore data at 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, especially those 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 industrialized data science pipelines See how commercial data scientists design scalable code and reusable code for data science services Explore cutting edge data science methods so that you can study trends and causality Discover advanced programming techniques using RDD and the DataFrame and Dataset APIs Find out how Spark can be used 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 Study advanced Spark concepts, solution design patterns, and integration architectures Demonstrate powerful data science pipelines In Detail Data science seeks to transform the world using data, and this is typically achieved through disrupting and changing real processes in real industries. In order to operate at this level you need to build data science solutions of substance –solutions that solve real problems. Spark has emerged as the big data platform of choice for data scientists due to its speed, scalability, and easy-to-use APIs. This book deep dives into using Spark to deliver production-grade data science solutions. This process is demonstrated by exploring the construction of a sophisticated global news analysis service that uses Spark to generate continuous geopolitical and current affairs insights. You will learn all about the core Spark APIs and take a comprehensive tour of advanced libraries, including Spark SQL, Spark Streaming, MLlib, and more. You will be introduced to advanced techniques and methods that will help you to construct commercial-grade data products. Focusing on a sequence of tutorials that deliver a working news intelligence service, you will learn about advanced Spark architectures, how to work with geographic data in Spark, and how to tune Spark algorithms so they scale linearly. Style and approach This is an advanced guide for those with beginner-level familiarity with the Spark architecture and working with Data Science applications. Mastering Spark for Data Science is a practical tutorial that uses core Spark APIs and takes a deep dive into advanced libraries including: Spark SQL, visual streaming, and MLlib. This book expands on titles like: Machine Learning with Spark and Learning Spark. It is the next learning curve for those comfortable with Spark and looking to improve their skills. 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.

Learn the basics of analytics on big data using Java, machine learning and other big data tools About This Book Acquire real-world set of tools for building enterprise level data science applications Surpasses the barrier of other languages in data science and learn create useful object-oriented codes Extensive use of Java compliant big data tools like apache spark, Hadoop, etc. Who This Book Is For This book is for Java developers who are looking to perform data analysis in production environment. Those who wish to implement data analysis in their Big data applications will find this book helpful. What You Will Learn Start from simple analytic tasks on big data Get into more complex tasks with predictive analytics on big data using machine learning Learn real time analytic tasks Understand the concepts with examples and case studies Prepare and refine data for analysis Create charts in order to understand the data See various real-world datasets In Detail This book covers case studies such as sentiment analysis on a tweet dataset, recommendations on a movielens dataset, customer segmentation on an ecommerce dataset, and graph analysis on actual flights dataset. This book is an end-to-end guide to implement analytics on big data with Java. Java is the de facto language for major big data environments, including Hadoop. This book will teach you how to perform analytics on big data with production-friendly Java. This book basically divided into two sections. The first part is an introduction that will help the readers get acquainted with big data environments, whereas the second part will contain a hardcore discussion on all the concepts in analytics on big data. It will take you from data analysis and data visualization to the core concepts and advantages of machine learning, real-life usage of regression and classification using Naive Bayes, a deep discussion on the concepts of clustering, and a review of simple neural networks on big data using deepLearning4j or plain Java Spark code. This book is a must-have book for Java developers who want to start learning big data analytics and want to use it in the real world. Style and approach The approach of book is to deliver practical learning modules in manageable content. Each chapter is a self-contained unit of a concept in big data analytics. Book will step by step builds the competency in the area of big data analytics. Examples using real world case studies to give ideas of real applications and how to use the techniques mentioned. The examples and case studies will be shown using both theory and code.

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

Summary Hadoop in Practice, Second Edition provides over 100 tested, instantly useful techniques that will help you conquer big data, using Hadoop. This revised new edition covers changes and new features in the Hadoop core architecture, including MapReduce 2. Brand new chapters cover YARN and integrating Kafka, Impala, and Spark SQL with Hadoop. You'll also get new and updated techniques for Flume, Sqoop, and Mahout, all of which have seen major new versions recently. In short, this is the most practical, up-to-date coverage of Hadoop available anywhere. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Book It's always a good time to upgrade your Hadoop skills! Hadoop in Practice, Second Edition provides a collection of 104 tested, instantly useful techniques for analyzing real-time streams, moving data securely, machine learning, managing large-scale clusters, and taming big data using Hadoop. This completely revised edition covers changes and new features in Hadoop core, including MapReduce 2 and YARN. You'll pick up hands-on best practices for integrating Spark, Kafka, and Impala with Hadoop, and get new and updated techniques for the latest versions of Flume, Sqoop, and Mahout. In short, this is the most practical, up-to-date coverage of Hadoop available. Readers need to know a programming language like Java and have basic familiarity with Hadoop. What's Inside Thoroughly updated for Hadoop 2 How to write YARN applications Integrate real-time technologies like Storm, Impala, and Spark Predictive analytics using Mahout and RR Readers need to know a programming language like Java and have basic familiarity with Hadoop. About the Author Alex Holmes works on tough big-data problems. He is a software engineer, author, speaker, and blogger specializing in large-scale Hadoop projects. Table of Contents PART 1 BACKGROUND AND FUNDAMENTALS Hadoop in a heartbeat Introduction to YARN PART 2 DATA LOGISTICS Data serialization—working with text and beyond Organizing and optimizing data in HDFS Moving data into and out of Hadoop PART 3 BIG DATA PATTERNS Applying MapReduce patterns to big data Utilizing data structures and algorithms at scale Tuning, debugging, and testing PART 4 BEYOND MAPREDUCE SQL on Hadoop Writing a YARN application Discover how graph algorithms can help you leverage the relationships within your data to develop more intelligent solutions and enhance your machine learning models. You'll learn how graph analytics are uniquely suited to unfold complex structures and reveal difficult-to-find patterns lurking in your data. Whether you are trying to build dynamic network models or forecast real-world behavior, this book illustrates how graph algorithms deliver value—from finding vulnerabilities and bottlenecks to detecting communities and improving machine learning predictions. This practical book walks you through hands-on examples of how to use graph algorithms in Apache Spark and Neo4j—two of the most common choices for graph analytics. Also included: sample code and tips for over 20 practical graph algorithms that cover optimal pathfinding, importance through centrality, and community detection. Learn how graph analytics vary from conventional statistical analysis Understand how classic graph algorithms work, and how they are applied Get guidance on which algorithms to use for different types of questions Explore algorithm examples with working code and sample datasets from Spark and Neo4j See how connected feature extraction can increase machine learning accuracy and precision Walk through creating an ML workflow for link prediction combining Neo4j and Spark Statistical Computation for Programmers, Scientists, Quants, Excel Users, and Other Professionals Using the open source R

language, you can build powerful statistical models to answer many of your most challenging questions. R has traditionally been difficult for non-statisticians to learn, and most R books assume far too much knowledge to be of help. *R for Everyone, Second Edition*, is the solution. Drawing on his unsurpassed experience teaching new users, professional data scientist Jared P. Lander has written the perfect tutorial for anyone new to statistical programming and modeling. Organized to make learning easy and intuitive, this guide focuses on the 20 percent of R functionality you'll need to accomplish 80 percent of modern data tasks. Lander's self-contained chapters start with the absolute basics, offering extensive hands-on practice and sample code. You'll download and install R; navigate and use the R environment; master basic program control, data import, manipulation, and visualization; and walk through several essential tests. Then, building on this foundation, you'll construct several complete models, both linear and nonlinear, and use some data mining techniques. After all this you'll make your code reproducible with LaTeX, RMarkdown, and Shiny. By the time you're done, you won't just know how to write R programs, you'll be ready to tackle the statistical problems you care about most. Coverage includes Explore R, RStudio, and R packages Use R for math: variable types, vectors, calling functions, and more Exploit data structures, including data.frames, matrices, and lists Read many different types of data Create attractive, intuitive statistical graphics Write user-defined functions Control program flow with if, ifelse, and complex checks Improve program efficiency with group manipulations Combine and reshape multiple datasets Manipulate strings using R's facilities and regular expressions Create normal, binomial, and Poisson probability distributions Build linear, generalized linear, and nonlinear models Program basic statistics: mean, standard deviation, and t-tests Train machine learning models Assess the quality of models and variable selection Prevent overfitting and perform variable selection, using the Elastic Net and Bayesian methods Analyze univariate and multivariate time series data Group data via K-means and hierarchical clustering Prepare reports, slideshows, and web pages with knitr Display interactive data with RMarkdown and htmlwidgets Implement dashboards with Shiny Build reusable R packages with devtools and Rcpp Register your product at informit.com/register for convenient access to downloads, updates, and corrections as they become available.

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.

Apache Spark's speed, ease of use, sophisticated analytics, and multilanguage support makes practical knowledge of this cluster-computing framework a required skill for data engineers and data scientists. With this hands-on guide, anyone looking for an introduction to Spark will learn practical algorithms and examples using PySpark. In each chapter, author Mahmoud Parsian shows you how to solve a data problem with a set of Spark transformations and algorithms. You'll learn how to tackle problems involving ETL, design patterns, machine learning algorithms, data partitioning, and genomics analysis. Each detailed recipe includes PySpark algorithms using the PySpark driver and shell script. With this book, you will: Learn how to select Spark transformations for optimized solutions Explore powerful transformations and reductions including `reduceByKey()`, `combineByKey()`, and `mapPartitions()` Understand data partitioning for optimized queries Design machine learning algorithms including Naive Bayes, linear regression, and logistic regression Build and apply a model using PySpark design patterns Apply motif-finding algorithms to graph data Analyze graph data by using the GraphFrames API Apply PySpark algorithms to clinical and genomics data (such as DNA-Seq)

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

Summary The Spark distributed data processing platform provides an easy-to-implement tool for ingesting, streaming, and processing data from any source. In *Spark in Action, Second Edition*, you'll learn to take advantage of Spark's core features and incredible processing speed, with applications including real-time computation, delayed evaluation, and machine learning. Spark skills are a hot commodity in enterprises worldwide, and with Spark's powerful and flexible Java APIs, you can reap all the benefits without first learning Scala or Hadoop. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications.

About the technology Analyzing enterprise data starts by reading, filtering, and merging files and streams from many sources. The Spark data processing engine handles this varied volume like a champ, delivering speeds 100 times faster than Hadoop systems. Thanks to SQL support, an intuitive interface, and a straightforward multilanguage API, you can use Spark without learning a complex new ecosystem.

About the book *Spark in Action, Second Edition*, teaches you to create end-to-end analytics applications. In this entirely new book, you'll learn from interesting Java-based examples, including a complete data pipeline for processing NASA satellite data. And you'll discover Java, Python, and Scala code samples hosted on GitHub that you can explore and adapt, plus appendixes that give you a cheat sheet for installing tools and understanding Spark-specific terms.

What's inside Writing Spark applications in Java Spark application architecture Ingestion through files, databases, streaming, and Elasticsearch Querying distributed datasets with Spark SQL

About the reader This book does not assume previous experience with Spark, Scala, or Hadoop.

About the author Jean-Georges Perrin is an experienced data and

software architect. He is France's first IBM Champion and has been honored for 12 consecutive years.

Table of Contents

PART 1 - THE THEORY CRIPPLED BY AWESOME EXAMPLES

- 1 So, what is Spark, anyway?
- 2 Architecture and flow
- 3 The majestic role of the dataframe
- 4 Fundamentally lazy
- 5 Building a simple app for deployment
- 6 Deploying your simple app

PART 2 - INGESTION

- 7 Ingestion from files
- 8 Ingestion from databases
- 9 Advanced ingestion: finding data sources and building your own
- 10 Ingestion through structured streaming

PART 3 - TRANSFORMING YOUR DATA

- 11 Working with SQL
- 12 Transforming your data
- 13 Transforming entire documents
- 14 Extending transformations with user-defined functions
- 15 Aggregating your data

PART 4 - GOING FURTHER

- 16 Cache and checkpoint: Enhancing Spark's performances
- 17 Exporting data and building full data pipelines
- 18 Exploring deployment

No need to spend hours ploughing through endless data – let Spark, one of the fastest big data processing engines available, do the hard work for you.

Key Features

- Get up and running with Apache Spark and Python
- Integrate Spark with AWS for real-time analytics
- Apply processed data streams to machine learning APIs of Apache Spark

Book Description

Processing big data in real time is challenging due to scalability, information consistency, and fault-tolerance. This book teaches you how to use Spark to make your overall analytical workflow faster and more efficient. You'll explore all core concepts and tools within the Spark ecosystem, such as Spark Streaming, the Spark Streaming API, machine learning extension, and structured streaming. You'll begin by learning data processing fundamentals using Resilient Distributed Datasets (RDDs), SQL, Datasets, and Dataframes APIs. After grasping these fundamentals, you'll move on to using Spark Streaming APIs to consume data in real time from TCP sockets, and integrate Amazon Web Services (AWS) for stream consumption. By the end of this book, you'll not only have understood how to use machine learning extensions and structured streams but you'll also be able to apply Spark in your own upcoming big data projects.

What you will learn

- Write your own Python programs that can interact with Spark
- Implement data stream consumption using Apache Spark
- Recognize common operations in Spark to process known data streams
- Integrate Spark streaming with Amazon Web Services (AWS)
- Create a collaborative filtering model with the movielens dataset
- Apply processed data streams to Spark machine learning APIs

Who this book is for

Data Processing with Apache Spark is for you if you are a software engineer, architect, or IT professional who wants to explore distributed systems and big data analytics. Although you don't need any knowledge of Spark, prior experience of working with Python is recommended.

Unlock the complexities of machine learning algorithms in Spark to generate useful data insights through this data analysis tutorial

About This Book

- Process and analyze big data in a distributed and scalable way
- Write sophisticated Spark pipelines that incorporate elaborate extraction
- Build and use regression models to predict flight delays

Who This Book Is For

Are you a developer with a background in machine learning and statistics who is feeling limited by the current

slow and “small data” machine learning tools? Then this is the book for you! In this book, you will create scalable machine learning applications to power a modern data-driven business using Spark. We assume that you already know the machine learning concepts and algorithms and have Spark up and running (whether on a cluster or locally) and have a basic knowledge of the various libraries contained in Spark.

What You Will Learn

- Use Spark streams to cluster tweets online
- Run the PageRank algorithm to compute user influence
- Perform complex manipulation of DataFrames using Spark
- Define Spark pipelines to compose individual data transformations
- Utilize generated models for off-line/on-line prediction
- Transfer the learning from an ensemble to a simpler Neural Network
- Understand basic graph properties and important graph operations
- Use GraphFrames, an extension of DataFrames to graphs, to study graphs using an elegant query language
- Use K-means algorithm to cluster movie reviews dataset

In Detail

The purpose of machine learning is to build systems that learn from data. Being able to understand trends and patterns in complex data is critical to success; it is one of the key strategies to unlock growth in the challenging contemporary marketplace today. With the meteoric rise of machine learning, developers are now keen on finding out how can they make their Spark applications smarter. This book gives you access to transform data into actionable knowledge. The book commences by defining machine learning primitives by the MLlib and H2O libraries. You will learn how to use Binary classification to detect the Higgs Boson particle in the huge amount of data produced by CERN particle collider and classify daily health activities using ensemble Methods for Multi-Class Classification. Next, you will solve a typical regression problem involving flight delay predictions and write sophisticated Spark pipelines. You will analyze Twitter data with help of the doc2vec algorithm and K-means clustering. Finally, you will build different pattern mining models using MLlib, perform complex manipulation of DataFrames using Spark and Spark SQL, and deploy your app in a Spark streaming environment.

Style and approach

This book takes a practical approach to help you get to grips with using Spark for analytics and to implement machine learning algorithms. We'll teach you about advanced applications of machine learning through illustrative examples. These examples will equip you to harness the potential of machine learning, through Spark, in a variety of enterprise-grade systems.

Data is bigger, arrives faster, and comes in a variety of formats—and it all needs to be processed at scale for analytics or machine learning. But how can you process such varied workloads efficiently? Enter Apache Spark. Updated to include Spark 3.0, this second edition shows data engineers and data scientists why structure and unification in Spark matters. Specifically, this book explains how to perform simple and complex data analytics and employ machine learning algorithms. Through step-by-step walk-throughs, code snippets, and notebooks, you'll be able to: Learn Python, SQL, Scala, or Java high-level Structured APIs Understand Spark operations and SQL Engine Inspect, tune, and debug Spark

operations with Spark configurations and Spark UI Connect to data sources: JSON, Parquet, CSV, Avro, ORC, Hive, S3, or Kafka Perform analytics on batch and streaming data using Structured Streaming Build reliable data pipelines with open source Delta Lake and Spark Develop machine learning pipelines with MLlib and productionize models using MLflow

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.

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. You'll start with an introduction to Spark and its ecosystem, and then dive into patterns that apply common techniques. Discover which machine learning tools make sense for particular problems.

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.

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

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 Learn the right cutting-edge skills and knowledge to leverage Spark Streaming to implement a wide array of real-time, streaming applications. This book walks you through end-to-end real-time application development using real-world applications, data, and code. Taking an application-first approach, each chapter introduces use cases from a specific industry and uses publicly available datasets from that domain to unravel the intricacies of production-grade design and implementation. The domains covered in Pro Spark Streaming include social media, the sharing economy, finance, online advertising, telecommunication, and IoT. In the last few years, Spark has become synonymous with big data processing. DStreams enhance the underlying Spark processing engine to support streaming analysis with a novel micro-batch processing model. Pro Spark Streaming by Zubair Nabi will enable you to become a specialist of latency sensitive applications by leveraging the key features of DStreams, micro-batch processing, and functional programming. To this end, the book includes ready-to-deploy examples and actual code. Pro Spark Streaming will act as the bible of Spark Streaming. What You'll Learn Discover Spark Streaming application development and best practices Work with the low-level details of discretized streams Optimize production-grade deployments of Spark Streaming via configuration recipes and instrumentation using Graphite, collectd, and Nagios Ingest data from disparate sources including MQTT, Flume, Kafka, Twitter, and a custom HTTP receiver Integrate and couple with HBase, Cassandra, and Redis Take advantage of design patterns for side-effects and maintaining state across the Spark Streaming micro-batch model Implement real-time and scalable ETL using data frames, SparkSQL, Hive, and SparkR Use streaming machine learning, predictive analytics, and recommendations Mesh batch processing with stream processing via the Lambda architecture Who This Book Is For Data scientists, big data experts, BI analysts, and data architects.

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

[Copyright: 42e62157f8b1ee4a5cda8f4309c72b8a](#)