Mahout In Action

Big Data Analytics with R and Hadoop is a tutorial style book that focuses on all the powerful big data tasks that can be achieved by integrating R and Hadoop.This book is ideal for R developers who are looking for a way to perform big data analytics with Hadoop. This book is also aimed at those who know Hadoop and want to build some intelligent applications over Big data with R packages. It would be helpful if readers have basic knowledge of R.

Written by renowned data science experts Foster Provost and Tom Fawcett, Data Science for Business introduces the fundamental principles of data science, and walks you through the "data-analytic thinking" necessary for extracting useful knowledge and business value from the data you collect. This guide also helps you understand the many data-mining techniques in use today. Based on an MBA course Provost has taught at New York University over the past ten years, Data Science for Business provides examples of real-world business problems to illustrate these principles. You'll not only learn how to improve communication between business stakeholders and data scientists, but also how participate intelligently in your company's data science projects. You'll also discover how to think data-analytically, and fully appreciate how data science methods can support business decision-making. Understand how data science fits in your organization—and how you can use it for competitive advantage Treat data as a business asset that requires careful investment if you're to gain real value Approach business problems data-analytically, using the data-mining process to gather good data in the most appropriate way Learn general concepts for actually extracting knowledge from data Apply data science principles when interviewing data science job candidates

Building a simple but powerful recommendation system is much easier than you think. Approachable for all levels of expertise, this report explains innovations that make machine learning practical for business production settings—and demonstrates how even a small-scale development team can design an effective large-scale recommendation system. Apache Mahout committers Ted Dunning and Ellen Friedman walk you through a design that relies on careful simplification. You'll learn how to collect the right data, analyze it with an algorithm from the Mahout library, and then easily deploy the recommender using search technology, such as Apache Solr or Elasticsearch. Powerful and effective, this efficient combination does learning offline and delivers rapid response recommendations in real time. Understand the tradeoffs between simple and complex recommenders Collect user data that tracks user actions—rather than their ratings Predict what a user wants based on behavior by others, using Mahoutfor co-occurrence analysis Use search technology to offer recommendations in real time, complete with item metadata Watch the recommender in action with a music service example Improve your recommender with dithering, multimodal recommendation, and other techniques

This book comprehensively covers the topic of recommender systems, which provide personalized recommendations of products or services to users based on their previous searches or purchases. Recommender system methods have been adapted to diverse applications including query log mining, social networking, news recommendations, and computational advertising. This book synthesizes both fundamental and advanced topics of a research area that has now reached maturity. The chapters of this book are organized into three categories: Algorithms and evaluation: These chapters discuss the fundamental algorithms in recommender systems, including collaborative filtering methods, content-based methods, knowledge-based methods, ensemble-based methods, and evaluation. Recommendations in specific domains and contexts: the context of a recommendation can be viewed as important side information that affects the recommendation goals. Different types of context such as temporal data, spatial data, social data, tagging data, and trustworthiness are explored. Advanced topics and applications: Various robustness aspects of recommender systems, such as shilling systems, attack models, and their defenses are discussed. In addition, recent topics, such as learning to rank, multi-armed bandits, group systems, multi-criteria systems, and active learning systems, are introduced together with applications. Although this book primarily serves as a textbook, it will also appeal to industrial practitioners and researchers due to its focus on applications and references. Numerous examples and exercises have been provided, and a solution manual is available for instructors.

Summary Liferay in Action is a comprehensive and authoritative guide to building portals on the Liferay 6 platform. Fully supported and authorized by Liferay, this book guides you smoothly from your first exposure to Liferay through the crucial day-to-day tasks of building and maintaining an enterprise portal that works well within your existing IT infrastructure. About the Technology A portal is a website built around a collection of components that request, display, and share information. Liferay Portal 6, an enterprise-ready development platform, makes it a snap to build portals that integrate with your existing backend systems and provide a rich interactive user experience. Because Liferay uses standard Java and JavaScript, along with built-in SOAP and JSON support for web services, developers can be productive immediately. And since it's available in both a free, open source version as well as a fully-supported commercial edition, it's an affordable solution for almost any business or organization About the Book Liferay in Action is the official guide to building Liferay portal applications using Java and JavaScript. If you've never used Liferay before, don't worry. This book starts with the basics: setting up your development environment and creating a working portal. Then, it builds on that foundation to help you discover social features, tagging, ratings, and more. You'll also explore the Portlet 2.0 API, and learn to create custom themes and reusable templates. Experienced developers will learn how to use new Liferay APIs to build social and collaborative sites, use the message bus and workflow, implement indexing and search, and more. This book was developed in close collaboration with Liferay engineers, so it answers the right questions, and answers them in depth. No experience with Liferay or the Portlets API is required, but basic knowledge of Java and web technology is assumed. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book. What's Inside Complete coverage of Liferay Portal 6 Covers both the commercial and open source versions Custom portlet development using the Portlet 2.0 spec Liferay's social network API Add functionality with hooks and Ext plugins ========== Table of Contents PART 1 WORKING WITH LIFERAY AND PORTLETS The Liferay difference Getting started with the Liferay development platform PART 2 WRITING APPLICATIONS ON LIFERAY'S PLATFORM A data-driven portlet made easy MVC the Liferay way Designing your site with themes and layout templates Making your site social Enabling user collaboration PART 3 CUSTOMIZING LIFERAY Hooks Extending Liferay effectively A tour of Liferay APIs

OpenCL in Action

The Official Guide to Liferay Portal Development

Liferay in Action

Apache Mahout

Collective Intelligence in Action

Hadoop in Action

Summary HBase in Action has all the knowledge you need to design, build, and run applications using HBase. First, it introduces you to the fundamentals of distributed systems and large scale data handling. Then, you'll explore real-world applications and code samples with just enough theory to understand the practical techniques. You'll see how to build applications with HBase and take advantage of the MapReduce processing framework. And along the way you'll learn patterns and best practices. About the Technology HBase is a NoSQL storage system designed for fast, random access to large volumes of data. It runs on commodity hardware and scales smoothly from modest datasets to billions of rows and millions of columns. About this Book HBase in Action is an experience-driven guide that shows you how to design, build, and run applications using HBase. First, it introduces you to the fundamentals of handling big data. Then, you'll explore HBase with the help of real applications and code samples and with just enough theory to back up the practical techniques. You'll take advantage of the MapReduce processing framework and benefit from seeing HBase best practices in action. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book. What's Inside When and how to use HBase Practical examples Design patterns for scalable data systems Deployment, integration, and design Written for developers and architects familiar with data storage and processing. No prior knowledge of HBase, Hadoop, or MapReduce is required. Table of Contents PART 1 HBASE FUNDAMENTALS Introducing HBase Getting started Distributed HBase, HDFS, and MapReduce PART 2 ADVANCED CONCEPTS HBase table design Extending HBase with coprocessors Alternative HBase clients PART 3 EXAMPLE APPLICATIONS HBase by example: OpenTSDB Scaling GIS on HBase PART 4 OPERATIONALIZING HBASE Deploying HBase Operations

This integrated collection covers a range of parallelization platforms, concurrent programming frameworks and machine learning settings, with case studies.

Apache Mahout Cookbook uses over 35 recipes packed with illustrations and real-world examples to help beginners as well as advanced programmers get acquainted with the features of Mahout."Apache Mahout Cookbook" is great for developers who want to have a fresh and fast introduction to Mahout coding. No previous knowledge of Mahout is required, and even skilled developers or system administrators will benefit from the various recipes presented

Apache Mahout: Beyond MapReduce. Distributed algorithm design This book is about designing mathematical and Machine Learning algorithms using the Apache Mahout "Samsara" platform. The material takes on best programming practices as well as conceptual approaches to attacking Machine Learning problems in big datasets. Math is explained, followed by code examples of distributed and in-memory computations. Written by Apache Mahout committers for people who want to learn how to design distributed math algorithms as well as how to use some of the new Mahout "Samsara" algorithms off-the-shelf. The book covers Apache Mahout 0.10 and 0.11.

Follows the author's apprenticeship to one of India's master conjurers and his initiation into the brotherhood of godmen, during which he journeys the subcontinent, meeting a plethora of adhus, sages, sorcerers, hypnotists, and humbugs. 15,000 first printing.

Giants of the Monsoon Forest: Living and Working with Elephants

Machine Learning and Big Data

Recipes for Scaling Up with Hadoop and Spark

What You Need to Know about Data Mining and Data-Analytic Thinking

Data Algorithms

Scaling Up Machine Learning

When Lucene first hit the scene five years ago, it was nothing short ofamazing. By using this open-source, highly scalable, super-fast search engine,developers could integrate search into applications quickly and efficiently.A lot has changed since then-search has grown from a "nice-to-have" featureinto an indispensable part of most enterprise applications. Lucene now powerssearch in diverse companies including Akamai, Netflix, LinkedIn,Technorati, HotJobs, Epiphany, FedEx, Mayo Clinic, MIT, New ScientistMagazine, and many others. Some things remain the same, though. Lucene still delivers high-performancesearch features in a disarmingly easy-to-use API. Due to its vibrant and diverseopen-source community of developers and users, Lucene is relentlessly improving,with evolutions to APIs, significant new features such as payloads, and ahuge increase (as much as 8x) in indexing speed with Lucene 2.3. And with clear writing, reusable examples, and unmatched advice on bestpractices, Lucene in Action, Second Edition is still the definitive guide todeveloping with Lucene. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book.

Summary OAuth 2 in Action teaches you the practical use and deployment of this HTTP-based protocol from the perspectives of a client, authorization server, and resource server. You'll learn how to confidently and securely build and deploy OAuth on both the client and server sides. Foreword by Ian Glazer. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Think of OAuth 2 as the web version of a valet key. It is an HTTP-based security protocol that allows users of a service to enable applications to use that service on their behalf without handing over full control. And OAuth is used everywhere, from Facebook and Google, to startups and cloud services. About the Book OAuth 2 in Action teaches you practical use and deployment of OAuth 2 from the perspectives of a client, an authorization server, and a resource server. You'll begin with an overview of OAuth and its components and interactions. Next, you'll get hands-on and build an OAuth client, an authorization server, and a protected resource. Then you'll dig into tokens, dynamic client registration, and more advanced topics. By the end, you'll be able to confidently and securely build and deploy OAuth on both the client and server sides. What's Inside Covers OAuth 2 protocol and design Authorization with OAuth 2 OpenID Connect and User-Managed Access Implementation risks JOSE, introspection, revocation, and registration Protecting and accessing REST APIs About the Reader Readers need basic programming skills and knowledge of HTTP and JSON. About the Author Justin Richer is a systems architect and software engineer. Antonio Sanso is a security software engineer and a security researcher. Both authors contribute to open standards and open source. Table of Contents Part 1 - First steps What is OAuth 2.0 and why should you care? The OAuth dance Part 2 - Building an OAuth 2 environment Building a simple OAuth client Building a simple OAuth protected resource Building a simple OAuth authorization server OAuth 2.0 in the real world Part 3 - OAuth 2 implementation and vulnerabilities Common client vulnerabilities Common protected resources vulnerabilities Common authorization server vulnerabilities Common OAuth token vulnerabilities Part 4 - Taking OAuth further OAuth tokens Dynamic client registration User authentication with OAuth 2.0 Protocols and profiles using OAuth 2.0 Beyond bearer tokens Summary and conclusions

The best-selling author of Big Data is back, this time with a unique and in-depth insight into how specific companies use big data. Big data is on the tip of everyone's tongue. Everyone understands its power and importance, but many fail to grasp the actionable steps and resources required to utilise it effectively. This book fills the knowledge gap by showing how major companies are using big data every day, from an up-close, on-the-ground perspective. From technology, media and retail, to sport teams, government agencies and financial institutions, learn the actual strategies and processes being used to learn about customers, improve manufacturing, spur innovation, improve safety and so much more. Organised for easy dip-in navigation, each chapter follows the same structure to give you the information you need quickly. For each company profiled, learn what data was used, what problem it solved and the processes put it place to make it practical, as well as the technical details, challenges and lessons learned from each unique scenario. Learn how predictive analytics helps Amazon, Target, John Deere and Apple understand their customers Discover how big data is behind the success of Walmart, LinkedIn, Microsoft and more Learn how big data is changing medicine, law enforcement, hospitality, fashion, science and banking Develop your own big data strategy by accessing additional reading materials at the end of each chapter

Hadoop in Action teaches readers how to use Hadoop and write MapReduce programs. The intended readers are programmers, architects, and project managers who have to process large amounts of data offline. Hadoop in Action will lead the reader from obtaining a copy of Hadoop to setting it up in a cluster and writing data analytic programs. The book begins by making the basic idea of Hadoop and MapReduce easier to grasp by applying the default Hadoop installation to a few easy-to-follow tasks, such as analyzing changes in word frequency across a body of documents. The book continues through the basic concepts of MapReduce applications developed using Hadoop, including a close look at framework components, use of Hadoop for a variety of data analysis tasks, and numerous examples of Hadoop in action. Hadoop in Action will explain how to use Hadoop and present design patterns and practices of programming MapReduce. MapReduce is a complex idea both conceptually and in its implementation, and Hadoop users are challenged to learn all the knobs and levers for running Hadoop. This book takes you beyond the mechanics of running Hadoop, teaching you to write meaningful programs in a MapReduce framework. This book assumes the reader will have a basic familiarity with Java, as most code examples will be written in Java. Familiarity with basic statistical concepts (e.g. histogram, correlation) will help the reader appreciate the more advanced data processing examples. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book.

#1 SAN FRANCISCO CHRONICLE BESTSELLER • NAMED ONE OF THE BEST BOOKS OF THE YEAR BY THE CHICAGO TRIBUNE "A picaresque, swashbuckling adventure."—The Washington Post Book World They're an odd pair, to be sure: pale, rail-thin, black-clad Zelikman, a moody, itinerant physician fond of jaunty headgear, and ex-soldier Amram, a gray-haired giant of a man as quick with a razor-tongued witticism as with a sharpened battle-ax. Brothers under the skin, comrades in arms, they make their rootless way through the Caucasus Mountains, circa a.d. 950, living as they please and surviving however they can—as blades and thieves for hire and as practiced bamboozlers, cheerfully separating the gullible from their money. But when they are dragooned into service as escorts and defenders to a prince of the Khazar Empire, they soon find themselves the half-willing generals in a full-scale revolution—on a road paved with warriors and whores, evil emperors and extraordinary elephants, secrets, swordplay, and such stuff as the grandest adventures are made of. Praise for Gentlemen of the Road "Within a few pages I was happily tangled in [Chabon's] net of finely filigreed language, seduced by an old-school-style swashbuckling quest . . . laced with surprises and humor."—San Francisco Chronicle "[Chabon] is probably the premiere prose stylist—the Updike—of his generation."—Time "The action is intricate and exuberant. . . . It's hard to resist its gathering momentum, not to mention the sheer headlong pleasure of Chabon's language."—The New York Times Book Review "[A] wild, wild adventure . . . abounds with lush language . . . This book roars to be read aloud."—Chicago Sun-Times

Concepts, Algorithms, Tools and Applications

Tusk and Stone

Elephant Run

Apache Oozie

Taming Text

Hadoop in Practice

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.

Summary Solr in Action is a comprehensive guide to implementing scalable search using Apache Solr. This clearly written book walks you through well-documented examples ranging from basic keyword searching to scaling a system for billions of documents and queries. It will give you a deep understanding of how to implement core Solr capabilities. About the Book Whether you're handling big (or small) data, managing documents, or building a website, it is important to be able to quickly search through your content and discover meaning in it. Apache Solr is your tool: a ready-to-deploy, Lucene-based, open source, full-text search engine. Solr can scale across many servers to enable real-time queries and data analytics across billions of documents. Solr in Action teaches you to implement scalable search using Apache Solr. This easy-to-read guide balances conceptual discussions with practical examples to show you how to implement all of Solr's core capabilities. You'll master topics like text analysis, faceted search, hit highlighting, result grouping, query suggestions, multilingual search, advanced geospatial and data operations, and relevancy tuning. This book assumes basic knowledge of Java and standard database technology. No prior knowledge of Solr or Lucene is required. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. What's Inside How to scale Solr for big data Rich real-world examples Solr as a NoSQL data store Advanced multilingual, data, and relevancy tricks Coverage of versions through Solr 4.7 About the Authors Trey Grainger is a director of engineering at CareerBuilder. Timothy Potter is a senior member of the engineering team at LucidWorks. The authors work on the scalability and reliability of Solr, as well as on recommendation engine and big data analytics technologies. Table of Contents PART 1 MEET SOLR Introduction to Solr Getting to know Solr Key Solr concepts Configuring Solr Indexing Text analysis PART 2 CORE SOLR CAPABILITIES Performing queries and handling results Faceted search Hit highlighting Query suggestions Result grouping/field collapsing Taking Solr to production PART 3 TAKING SOLR TO THE NEXT LEVEL SolrCloud Multilingual search Complex query operations Mastering relevancy

Summary Taming Text, winner of the 2013 Jolt Awards for Productivity, is a hands-on, example-driven guide to working with unstructured text in the context of real-world applications. This book explores how to automatically organize text using approaches such as full-text search, proper name recognition, clustering, tagging, information extraction, and summarization. The book guides you through examples illustrating each of these topics, as well as the foundations upon which they are built. About this Book There is so much text in our lives, we are practically drowningin it. Fortunately, there are innovative tools and techniquesfor managing unstructured information that can throw thesmart developer a much-needed lifeline. You'll find them in thisbook. Taming Text is a practical, example-driven guide to working withtext in real applications. This book introduces you to useful techniques like full-text search, proper name recognition,clustering, tagging, information extraction, and summarization.You'll explore real use cases as you systematically absorb thefoundations upon which they are built.Written in a clear and concise style, this book avoids jargon, explainingthe subject in terms you can understand without a backgroundin statistics or natural language processing. Examples arein Java, but the concepts can be applied in any language. Written for Java developers, the book requires no prior knowledge of GWT. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from

Manning. Also available is all code from the book. Winner of 2013 Jolt Awards: The Best Books—one of five notable books every serious programmer should read. What's Inside When to use text-taming techniques Important open-source libraries like Solr and Mahout How to build text-processing applications About the Authors Grant Ingersoll is an engineer, speaker, and trainer, a Lucenecommitter, and a cofounder of the Mahout machine-learning project. Thomas Morton is the primary developer of OpenNLP and Maximum Entropy. Drew Farris is a technology consultant, software developer, and contributor to Mahout,Lucene, and Solr. "Takes the mystery out of verycomplex processes."—From the Foreword by Liz Liddy, Dean, iSchool, Syracuse University Table of Contents Getting started taming text Foundations of taming text Searching Fuzzy string matching Identifying people, places, and things Clustering text Classification, categorization, and tagging Building an example question answering system Untamed text: exploring the next frontier

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

Until now, design patterns for the MapReduce framework have been scattered among various research papers, blogs, and books. This handy guide brings together a unique collection of valuable MapReduce patterns that will save you time and effort regardless of the domain, language, or development framework you're using. Each pattern is explained in context, with pitfalls and caveats clearly identified to help you avoid common design mistakes when modeling your big data architecture. This book also provides a complete overview of MapReduce that explains its origins and implementations, and why design patterns are so important. All code examples are written for Hadoop. Summarization patterns: get a top-level view by summarizing and grouping data Filtering patterns: view data subsets such as records generated from one user Data organization patterns: reorganize data to work with other systems, or to make MapReduce analysis easier Join patterns: analyze different datasets together to discover interesting relationships Metapatterns: piece together several patterns to solve multi-stage problems, or to perform several analytics in the same job Input and output patterns: customize the way you use Hadoop to load or store data "A clear exposition of MapReduce programs for common data processing patterns—this book is indespensible for anyone using Hadoop." --Tom White, author of Hadoop: The Definitive Guide

Gentlemen of the Road

HBase in Action

Big Data Analytics with R and Hadoop

Sorcerer's Apprentice

MapReduce Design Patterns

A Million Shades of Gray

Presents information on machine learning through the use of Apache Mahout, covering such topics as using group data to make individual recommendations, finding logical clusters, and filtering classifications.

Summary OpenCL in Action is a thorough, hands-on presentation of OpenCL, with an eye toward showing developers how to build high-performance applications of their own. It begins by presenting the core concepts behind OpenCL, including vector computing, parallel programming, and multi-threaded operations, and then guides you step-by-step from simple data structures to complex functions. About the Technology Whatever system you have, it probably has more raw processing power than you're using. OpenCL is a high-performance programming language that maximizes computational power by executing on CPUs, graphics processors, and other number-crunching devices. It's perfect for speed-sensitive tasks like vector computing, matrix operations, and graphics acceleration. About this Book OpenCL in Action blends the theory of parallel computing with the practical reality of building high-performance applications using OpenCL. It first guides you through the fundamental data structures in an intuitive manner. Then, it explains techniques for high-speed sorting, image processing, matrix operations, and fast Fourier transform. The book concludes with a deep look at the all-important subject of graphics acceleration. Numerous challenging examples give you different ways to experiment with working code. A background in C or C++ is helpful, but no prior exposure to OpenCL is needed. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book. What's Inside Learn OpenCL step by step Tons of annotated code Tested algorithms for maximum performance ******** Table of Contents PART 1 FOUNDATIONS OF OPENCL PROGRAMMING Introducing OpenCL Host programming: fundamental data structures Host programming: data transfer and partitioning Kernel programming: data types and device memory Kernel programming: operators and functions Image processing Events, profiling, and synchronization Development with C++ Development with Java and Python General coding principles PART 2 CODING PRACTICAL ALGORITHMS IN OPENCL Reduction and sorting Matrices and QR decomposition Sparse matrices Signal processing and the fast Fourier transform PART 3 ACCELERATING OPENGL WITH OPENCL Combining OpenCL and OpenGL Textures and renderbuffers**

Jess in Action first introduces rule programming concepts and teaches you the Jess language. Armed with this knowledge, you then progress through a series of fully-developed applications chosen to expose you to practical rule-based development. The book shows you how you can add power and intelligence to your Java software.

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

After Nick's British father's plantation in Burma is invaded by the Japanese in 1941, and his father and others are taken prisoner, Nick and his friend Mya plan a daring escape on elephants, risking their lives to save Nick's father and Mya's brother from the prisoner of war camp.

Big Data Analytics Beyond Hadoop

Building Effective Algorithms and Analytics for Hadoop and Other Systems

Data-intensive Text Processing with MapReduce

A Tale of Adventure

Machine Learning in Action

Beyond Mapreduce

Summary Tika in Action is a hands-on guide to content mining with Apache Tika. The book's many examples and case studies offer real-world experience from domains ranging from search engines to digital asset management and scientific data processing. About the Technology Tika is an Apache toolkit that has built into it everything you and your app need to know about file formats. Using Tika, your applications can discover and extract content from digital documents in almost any format, including exotic ones. About this Book Tika in Action is the ultimate guide to content mining using Apache Tika. You'll learn how to pull usable information from otherwise inaccessible sources, including internet media and file archives. This example-rich book teaches you to build and extend applications based on real-world experience with search engines, digital asset management, and scientific data processing. In addition to architectural overviews, you'll find detailed chapters on features like metadata extraction, automatic language detection, and custom parser development. This book is written for developers who are new to both Scala and Lift and covers just enough Scala to get you started. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book. What's Inside Crack MS Word, PDF, HTML, and ZIP Integrate with search engines, CMS, and other data sources Learn through experimentation Many examples This book requires no previous knowledge of Tika or text mining techniques. It assumes a working knowledge of Java. ==========?== Table of Contents PART 1 GETTING STARTED The case for the digital Babel fish Getting started with Tika The information landscape PART 2 TIKA IN DETAIL Document type detection Content extraction Understanding metadata Language detection What's in a file? PART 3 INTEGRATION AND ADVANCED USE The big picture Tika and the Lucene search stack Extending Tika PART 4 CASE STUDIES Powering NASA science data systems Content management with Apache Jackrabbit Curating cancer research data with Tika The classic search engine example

Our world is being revolutionized by data-driven methods: access to large amounts of data has generated new insights and opened exciting new opportunities in commerce, science, and computing applications. Processing the enormous quantities of data necessary for these advances requires large clusters, making distributed computing paradigms more crucial than ever. MapReduce is a programming model for expressing distributed computations on massive datasets and an execution framework for large-scale data processing on clusters of commodity servers. The programming model provides an easy-to-understand abstraction for designing scalable algorithms, while the execution framework transparently handles many system-level details, ranging from scheduling to synchronization to fault tolerance. This book focuses on MapReduce algorithm design, with an emphasis on text processing algorithms common in natural language processing, information retrieval, and machine learning. We introduce the notion of MapReduce design patterns, which represent general reusable solutions to commonly occurring problems across a variety of problem domains. This book not only intends to help the reader "think in MapReduce", but also discusses limitations of the programming model as well. This volume is a printed version of a work that appears in the Synthesis Digital Library of Engineering and Computer Science. Synthesis Lectures provide concise, original presentations of important research and development topics, published quickly, in digital and print formats. For more information visit www.morganclaypool.com

There's a great deal of wisdom in a crowd, but how do you listen to a thousand people talking at once? Identifying the wants, needs, and knowledge of internet users can be like listening to a mob. In the Web 2.0 era, leveraging the collective power of user contributions, interactions, and feedback is the key to market dominance. A new category of powerful programming techniques lets you discover the patterns, inter-relationships, and individual profiles-the collective intelligence--locked in the data people leave behind as they surf websites, post blogs, and interact with other users. Collective Intelligence in Action is a hands-on guidebook for implementing collective intelligence concepts using Java. It is the first Java-based book to emphasize the underlying algorithms and technical implementation of vital data gathering and mining techniques like analyzing trends, discovering relationships, and making predictions. It provides a pragmatic approach to personalization by combining content-based analysis with collaborative approaches. This book is for Java developers implementing Collective Intelligence in real, high-use applications. Following a running example in which you harvest and use information from blogs, you learn to develop software that you can embed in your own applications. The code examples are immediately reusable and give the Java developer a working collective intelligence toolkit. Along the way, you work with, a number of APIs and open-source toolkits including text analysis and search using Lucene, web-crawling using Nutch, and applying machine learning algorithms using WEKA and the Java Data Mining (JDM) standard. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book.

Mahout in ActionManning Publications

Summary RabbitMQ in Depth is a practical guide to building and maintaining message-based applications. This book provides detailed coverage of RabbitMQ with an emphasis on why it works the way it does. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology At the heart of most modern distributed applications is a queue that buffers, prioritizes, and routes message traffic. RabbitMQ is a high-performance message broker based on the Advanced Message Queueing Protocol. It?s battle tested, ultrafast, and powerful enough to handle anything you can throw at it. It requires a few simple setup steps, and you can instantly start using it to manage low-level service communication, application integration, and distributed system message routing. About the Book RabbitMQ in Depth is a practical guide to building and maintaining message-based applications. This book provides detailed coverage of RabbitMQ with an emphasis on why it works the way it does. You'll find examples and detailed explanations based in real-world systems ranging from simple networked services to complex distributed designs. You'll also find the insights you need to make core architectural choices and develop procedures for effective operational management. What's Inside AMQP, the Advanced Message Queueing Protocol Communicating via MQTT, Stomp, and HTTP Valuable troubleshooting techniques Database integration About the Reader Written for programmers with a basic understanding of messaging-oriented systems. About the Author Gavin M. Roy is an active, open source evangelist and advocate who has been working with internet and enterprise technologies since the mid-90s. Technical editor James Titcumb is a freelance developer, trainer, speaker, and active contributor to open source projects. Table of Contents PART 1 - RABBITMQ AND APPLICATION ARCHITECTURE Foundational RabbitMQ How to speak Rabbit: the AMQ Protocol An in-depth tour of message properties Performance trade-offs in publishing Don't get messages; consume them Message patterns via exchange routing PART 2 - MANAGING RABBITMQ IN THE DATA CENTER OR THE CLOUD Scaling RabbitMQ with clusters Cross-cluster message distribution PART 3 - INTEGRATIONS AND CUSTOMIZATION Using alternative protocols Database integrations

Practical Machine Learning

Recommender Systems

How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results

Lucene in Action

RabbitMQ in Depth

How to accelerate graphics and computations

After a criminal gang attacks his caravan and he loses his identity as a Brahmin, Arjun resigns himself to his new life as a soldier, becomes an elephant driver, and searches for his kidnapped sister.

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

Summary Machine Learning in Action is unique book that blends the foundational theories of machine learning with the practical realities of building tools for everyday data analysis. You'll use the flexible Python programming language to build programs that implement algorithms for data classification, forecasting, recommendations, and higher-level features like summarization and simplification. About the Book A machine is said to learn when its performance improves with experience. Learning requires algorithms and programs that capture data and ferret out the interestingor useful patterns. Once the specialized domain of analysts and mathematicians, machine learning is becoming a skill needed by many. Machine Learning in Action is a clearly written tutorial for developers. It avoids academic language and takes you straight to the techniques you'll use in your day-to-day work. Many (Python) examples present the core algorithms of statistical data processing, data analysis, and data visualization in code you can reuse. You'll understand the concepts and how they fit in with tactical tasks like classification, forecasting, recommendations, and higher-level features like summarization and simplification. Readers need no prior experience with machine learning or statistical processing. Familiarity with Python is helpful. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book. What's Inside A no-nonsense introduction Examples showing common ML tasks Everyday data analysis Implementing classic algorithms like Apriori and Adaboos Table of Contents PART 1 CLASSIFICATION Machine learning basics Classifying with k-Nearest Neighbors Splitting datasets one feature at a time: decision trees Classifying with probability theory: na ï ve Bayes Logistic regression Support vector machines Improving classification with the AdaBoost meta algorithm PART 2 FORECASTING NUMERIC VALUES WITH REGRESSION Predicting numeric values: regression Tree-based regression PART 3 UNSUPERVISED LEARNING Grouping unlabeled items using k-means clustering Association analysis with the Apriori algorithm Efficiently finding frequent itemsets with FP-growth PART 4 ADDITIONAL TOOLS Using principal component analysis to simplify data Simplifying data with the singular value decomposition Big data and MapReduce

Summary Spark in Action teaches you the theory and skills you need to effectively handle batch and streaming data using Spark. Fully updated for Spark 2.0. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Big data systems distribute datasets across clusters of machines, making it a challenge to efficiently query, stream, and interpret them. Spark can help. It is a processing system designed specifically for distributed data. It provides easy-to-use interfaces, along with the performance you need for production-quality analytics and machine learning. Spark 2 also adds improved programming APIs, better performance, and countless other upgrades. About the Book Spark in Action teaches you the theory and skills you need to effectively handle batch and streaming data using Spark. You'll get comfortable with the Spark CLI as you work through a few introductory examples. Then, you'll start programming Spark using its core APIs. Along the way, you'll work with structured data using Spark SQL, process near-real-time streaming data, apply machine learning algorithms, and munge graph data using Spark GraphX. For a zero-effort startup, you can download the preconfigured virtual machine ready for you to try the book's code. What's Inside Updated for Spark 2.0 Real-life case studies Spark DevOps with Docker Examples in Scala, and online in Java and Python About the Reader Written for experienced programmers with some background in big data or machine learning. About the Authors Petar Ze č evi ć and Marko Bona ć i are seasoned developers heavily involved in the Spark community. Table of Contents PART 1 - FIRST STEPS Introduction to Apache Spark Spark fundamentals Writing Spark applications The Spark API in depth PART 2 - MEET THE SPARK FAMILY Sparkling queries with Spark SQL Ingesting data with Spark Streaming Getting smart with MLlib ML: classification and clustering Connecting the dots with GraphX PART 3 - SPARK OPS Running Spark Running on a Spark standalone cluster Running on YARN and Mesos PART 4 - BRINGING IT TOGETHER Case study: real-time dashboard Deep learning on Spark with H2O

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)

Spark in Action

Real-Time Applications with Storm, Spark, and More Hadoop Alternatives

The Workflow Scheduler for Hadoop

Solr in Action

Data Science for Business

How to Find, Organize, and Manipulate It

Get a solid grounding in Apache Oozie, the workflow scheduler system for managing Hadoop jobs. With this hands-on guide, two experienced Hadoop practitioners walk you through the intricacies of this powerful and flexible platform, with numerous examples and real-world use cases. Once you set up your Oozie server, you'll dive into techniques for writing and coordinating workflows, and learn how to write complex data pipelines. Advanced topics show you how to handle shared libraries in Oozie, as well as how to implement and manage Oozie's security capabilities. Install and configure an Oozie server, and get an overview of basic concepts Journey through the world of writing and configuring workflows Learn how the Oozie coordinator schedules and executes workflows based on triggers Understand how Oozie manages data dependencies Use Oozie bundles to package several coordinator apps into a data pipeline Learn about security features and shared library management Implement custom extensions and write your own EL functions and actions Debug workflows and manage Oozie's operational details

Currently many different application areas for Big Data (BD) and Machine Learning (ML) are being explored. These promising application areas for BD/ML are the social sites, search engines, multimedia sharing sites, various stock exchange sites, online gaming, online survey sites and various news sites, and so on. To date, various use-cases for this application area are being researched and developed. Software applications are already being published and used in various settings from education and training to discover useful hidden patterns and other information like customer choices and market trends that can help organizations make more informed and customer-oriented business decisions. Combining BD with ML will provide powerful, largely unexplored application areas that will revolutionize practice in Videos Surveillance, Social Media Services, Email Spam and Malware Filtering, Online Fraud Detection, and so on. It is very important to continuously monitor and understand these effects from safety and societal point of view. Hence, the main purpose of this book is for researchers, software developers and practitioners, academicians and students to showcase novel use-cases and applications, present empirical research results from user-centered qualitative and quantitative experiments of these new applications, and facilitate a discussion forum to explore the latest trends in big data and machine learning by providing algorithms which can be trained to perform interdisciplinary techniques such as statistics, linear algebra, and optimization and also create automated systems that can sift through large volumes of data at high speed to make predictions or decisions without human intervention

A boy and his elephant escape into the jungle when the Viet Cong attack his village immediately after the Vietnam war.

Master alternative Big Data technologies that can do what Hadoop can't: real-time analytics and iterative machine learning. When most technical professionals think of Big Data analytics today, they think of Hadoop. But there are many cutting-edge applications that Hadoop isn't well suited for, especially real-time analytics and contexts requiring the use of iterative machine learning algorithms. Fortunately, several powerful new technologies have been developed specifically for use cases such as these. Big Data Analytics Beyond Hadoop is the first guide specifically designed to help you take the next steps beyond Hadoop. Dr. Vijay Srinivas Agneeswaran introduces the breakthrough Berkeley Data Analysis Stack (BDAS) in detail, including its motivation, design, architecture, Mesos cluster management, performance, and more. He presents realistic use cases and up-to-date example code for: Spark, the next generation in-memory computing technology from UC Berkeley Storm, the parallel real-time Big Data analytics technology from Twitter GraphLab, the next-generation graph processing paradigm from CMU and the University of Washington (with comparisons to alternatives such as Pregel and Piccolo) Halo also offers architectural and design guidance and code sketches for scaling machine learning algorithms to Big Data, and then realizing them in real-time. He concludes by previewing emerging trends, including real-time video analytics, SDNs, and even Big Data governance, security, and privacy issues. He identifies intriguing startups and new research possibilities, including BDAS extensions and cutting-edge model-driven analytics. Big Data Analytics Beyond Hadoop is an indispensable resource for everyone who wants to reach the cutting edge of Big Data analytics, and stay there: practitioners, architects, programmers, data scientists, researchers, startup entrepreneurs, and advanced students.

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.

OAuth 2 in Action

Big Data in Practice

The Textbook

Apache Mahout Cookbook

Mahout in Action

Practical Machine Learning: Innovations in Recommendation

"No one who loves elephants or how humans interact with wildlife should pass up Jacob Shell's remarkable book." —Dan Flores, author of Coyote America Giants of the Monsoon Forest journeys deep into the mountainous rainforests of Burma and India to explore the world of teak logging elephants and their intriguing alliance with humans. Jacob Shell's narrative vividly depicts elephants' extraordinary intelligence, and the complicated bond with individual human riders, a partnership that can last for decades. Giants of the Monsoon Forest reveals an unexpected relationship between evolution in the natural world and political struggles in the human one, while considering how Asia's secret forest culture might offer a way to help protect the fragile spaces both elephants and humans need to survive.

Rule-Based Systems in Java

Tika in Action

Parallel and Distributed Approaches

An Introduction for Data Scientists

Data Analytics with Hadoop