Hadoop In Practice Alex Holmes

Yeah, reviewing a books **Hadoop In Practice Alex Holmes** could mount up your close contacts listings. This is just one of the solutions for you to be successful. As understood, capability does not suggest that you have wonderful points.

Comprehending as with ease as treaty even more than additional will come up with the money for each success. next to, the proclamation as well as sharpness of this Hadoop In Practice Alex Holmes can be taken as well as picked to act.

Hadoop For Dummies - Dirk deRoos 2014-04-14
Let Hadoop For Dummies help harness the power of your data and rein in the information overload Big data has become big business, and companies and organizations of all sizes are struggling to find ways to retrieve valuable information from their massive data sets with becoming overwhelmed. Enter Hadoop and this easy-to-understand For Dummies guide. Hadoop For Dummies helps readers understand the value of big data, make a business case for using Hadoop, navigate the Hadoop ecosystem, and build and manage Hadoop applications and clusters. Explains the origins of Hadoop, its economic benefits, and its functionality and practical applications Helps you find your way around the Hadoop ecosystem, program MapReduce, utilize design patterns, and get your Hadoop cluster up and running quickly and easily Details how to use Hadoop applications for data mining, web analytics and personalization, large-scale text processing, data science, and problem-solving Shows you how to improve the value of your Hadoop cluster, maximize your investment in Hadoop, and avoid common pitfalls when building your Hadoop cluster From programmers challenged with building and maintaining affordable, scaleable data systems to administrators who must deal with huge volumes of information effectively and efficiently, this how-to has something to help you with Hadoop.

Data-intensive Text Processing with MapReduce - Jimmy Lin 2010
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

Hadoop in Practice - Alex Holmes 2014-10-12
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

Forest-Water Interactions - Delphis F. Levia 2020-02-05
The United Nations has declared 2018-2028 as the International Decade for Action on Water for Sustainable Development. This is a timely designation. In an increasingly thirsty world, the subject of forest-water interactions is of critical importance to the achievement of sustainability goals. The central underlying tenet of this book is that the hydrologic community can conduct better science and make a more meaningful impact to the world’s water crisis if scientists are: (1) better equipped to utilize new methods and harness big data from either or both high-frequency sensors and long-term research watersheds; and (2) aware of new developments in our process-based understanding of the hydrological cycle in both natural and urban settings. Accordingly, this forward-looking book delves into forest-water interactions from multiple methodological, statistical, and process-based perspectives (with some chapters featuring data sets and open-source R code), concluding with a chapter on future forest hydrology under global change. Thus, this book describes the opportunities of convergence in high-frequency sensing, big data, and open source software to catalyze more comprehensive understanding of forest-water interactions. The book will be of interest to researchers, graduate students, and advanced undergraduates in an array of disciplines, including hydrology, forestry, ecology, botany, and environmental engineering.

Big Data and Analytics - Vincenzo Morabito 2015-01-31
This book presents and discusses the main strategic and organizational challenges posed by Big Data and analytics in a manner relevant to both practitioners and scholars. The first part of the book analyzes strategic issues relating to the growing relevance of Big Data and analytics for competitive advantage, which is also attributable to empowerment of activities such as consumer profiling, market segmentation, and development of new products or services. Detailed consideration is also given to the strategic impact of Big Data and analytics on innovation in domains such as government and education and to Big Data-driven business models. The second part of the book addresses the impact of Big Data and analytics on

management and organizations, focusing on challenges for governance, evaluation, and change management, while the concluding part reviews real examples of Big Data and analytics innovation at the global level. The text is supported by informative illustrations and case studies, so that practitioners can use the book as a toolbox to improve understanding and exploit business opportunities related to Big Data and analytics.

Handbook of e-Business Security - João Manuel R.S. Tavares 2018-07-27

There are a lot of e-business security concerns. Knowing about e-business security issues will likely help overcome them. Keep in mind, companies that have control over their e-business are likely to prosper most. In other words, setting up and maintaining a secure e-business is essential and important to business growth. This book covers state-of-the art practices in e-business security, including privacy, trust, security of transactions, big data, cloud computing, social network, and distributed systems.

Programming Hive - Edward Capriolo 2012-09-26

Describes the features and functions of Apache Hive, the data infrastructure for Hadoop.

Data Algorithms - Mahmoud Parsian 2015-07-13

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

Professional Hadoop Solutions - Boris Lublinsky 2013-09-12

The go-to guidebook for deploying Big Data solutions withHadoop Today's enterprise architects need to understand how the Hadoopframeworks and APIs fit together, and how they can be integrated todeliver real-world solutions. This book is a practical, detailedguide to building and implementing those solutions, with code-levelinstruction in the popular Wrox tradition. It covers storing datawith HDFS and Hbase, processing data with MapReduce, and automatingdata processing with Oozie. Hadoop security, running Hadoop withAmazon Web Services, best practices, and automating Hadoopprocesses in real time are also covered in depth. With in-depth code examples in Java and XML and the latest onrecent additions to the Hadoop ecosystem, this complete resourcealso covers the use of APIs, exposing their inner workings andallowing architects and developers to better leverage and customizethem. The ultimate guide for developers, designers, and architectswho need to build and deploy Hadoop applications Covers storing and processing data with various technologies,automating data processing, Hadoop security, and deliveringreal-time solutions Includes detailed, real-world examples and code-levelguidelines Explains when, why, and how to use these tools effectively Written by a team of Hadoop experts in theprogrammer-to-programmer Wrox style Professional Hadoop Solutions is the reference enterprisearchitects and developers need to maximize the power of Hadoop.

Frontiers in Massive Data Analysis - National Research Council 2013-09-03

Data mining of massive data sets is transforming the way we think about crisis response, marketing, entertainment, cybersecurity and national intelligence. Collections of documents, images, videos, and networks are being thought of not merely as bit strings to be stored, indexed, and retrieved, but as potential sources of discovery and knowledge, requiring sophisticated analysis techniques that go far beyond classical indexing and keyword counting, aiming to find relational and semantic interpretations of the phenomena underlying the data. Frontiers in Massive Data Analysis examines the frontier of analyzing massive amounts of data, whether in a static database or streaming through a system. Data at that scale--terabytes and petabytes--is increasingly common in science (e.g., particle physics, remote sensing, genomics), Internet commerce, business analytics, national security, communications, and elsewhere. The tools that work to infer knowledge from data at smaller scales do not necessarily work, or work well, at such massive scale. New tools, skills, and approaches are necessary, and this report identifies many of them, plus promising research directions to explore. Frontiers in Massive Data Analysis discusses pitfalls in trying to infer knowledge from massive data, and it characterizes seven major classes of computation that are common in the analysis of massive data. Overall, this report illustrates the cross-disciplinary knowledge--from computer science, statistics, machine learning, and application disciplines--that must be brought to bear to make useful inferences from massive data.

Strategy Without Design - Robert C. H. Chia 2009-10-08

"In business the survival and flourishing of an organisation is most often associated with the ability of its strategists to create a distinctive identity by confronting and rising above others. Yet not all organisational accomplishment can be explained with recourse to deliberate choice and purposeful design on the part of strategic actors. This book shows why. Using examples from the world of business, economics, military strategy, politics and philosophy, it argues that collective success may inadvertently emerge as a result of the everyday coping actions of a multitude of individuals, none of whom intended to contribute to any preconceived plan. A consequence of this claim is that a paradox exists in strategic interventions, one that no strategist can afford to ignore. The more directly and deliberately a strategic goal is single-mindedly sought, the more likely it is that such calculated instrumental action eventually works to undermine its own initial success"--Provided by publisher.

Producing Open Source Software - Karl Fogel 2005-10-07

The corporate market is now embracing free, "open source" software like never before, as evidenced by the recent success of the technologies underlying LAMP (Linux, Apache, MySQL, and PHP). Each is the result of a publicly collaborative process among numerous developers who volunteer their time and energy to create better software. The truth is, however, that the overwhelming majority of free software projects fail. To help you beat the odds, O'Reilly has put together Producing Open Source Software, a guide that recommends tried and true steps to help free software developers work together toward a common goal. Not just for developers who are considering starting their own free software project, this book will also help those who want to participate in the process at any level. The book tackles this very complex topic by distilling it down into easily understandable parts. Starting with the basics of project management, it details specific tools used in free software projects, including version control, IRC, bug tracking, and Wikis. Author Karl Fogel, known for his work on CVS and Subversion, offers practical advice on how to set up and use a range of tools in combination with open mailing lists and archives. He also provides several chapters on the essentials of recruiting and motivating developers, as well as how to gain much-needed publicity for your project. While managing a team of enthusiastic developers -- most of whom you've never even met -- can be challenging, it can also be fun. Producing Open Source Software takes this into account, too, as it speaks of the sheer pleasure to be had from working with a motivated team of free software developers.

Introduction to Cryptography and Network Security - Behrouz A. Forouzan 2008

In this new first edition, well-known author Behrouz Forouzan uses his accessible writing style and visual approach to simplify the difficult concepts of cryptography and network security. While many security books assume knowledge of number theory and advanced math, or present mainly theoretical ideas, Forouzan presents difficult security topics from the ground up. A gentle introduction to the fundamentals of number theory is provided in the opening chapters, paving the way for the student to move on to more complex security and cryptography topics. Difficult math concepts are organized in appendices at the end of each chapter so that students can first learn the principles, then apply the technical background. Hundreds of examples, as well as fully coded programs, round out a practical, hands-on approach which encourages students to test the material they are learning.

Advanced Analytics with Spark - Sandy Ryza 2015-04-02

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

Just Enough Software Architecture - George Fairbanks 2010-08-30

This is a practical guide for software developers, and different than other software architecture books. Here's why: It teaches risk-driven architecting. There is no need for meticulous designs when risks are small, nor any excuse for sloppy designs when risks threaten your success. This book describes a way to do just enough architecture. It avoids the one-size-fits-all process tar pit with advice on how to tune your design effort based on the risks you face. It democratizes architecture. This book seeks to make architecture relevant to all software developers. Developers need to understand how to use constraints as guiderails that ensure desired outcomes, and how seemingly small changes can affect a system's properties. It cultivates declarative knowledge. There is a difference between being able to hit a ball and knowing why you are able to hit it, what psychologists refer to as procedural knowledge versus declarative knowledge. This book will make you more aware of what you have been doing and provide names for the concepts. It emphasizes the engineering. This book focuses on the technical parts of software development and what developers do to ensure the system works not job titles or processes. It shows you how to build models and analyze architectures so that you can make principled design tradeoffs. It describes the techniques software designers use to reason about medium to large sized problems and points out where you can learn specialized techniques in more detail. It provides practical advice. Software design decisions influence the architecture and vice versa. The approach in this book embraces drill-down/pop-up behavior by describing models that have various levels of abstraction, from architecture to data structure design.

Big Data - Nathan Marz 2015

Summary Big Data teaches you to build big data systems using an architecture that takes advantage of clustered hardware along with new tools designed specifically to capture and analyze web-scale data. It describes a scalable, easy-to-understand approach to big data systems that can be built and run by a small team. Following a realistic example, this book guides readers through the theory of big data systems, how to implement them in practice, and how to deploy and operate them once they're built. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Book Web-scale applications like social networks, real-time analytics, or e-commerce sites deal with a lot of data, whose volume and velocity exceed the limits of traditional database systems. These applications require architectures built around clusters of machines to store and process data of any size, or speed. Fortunately, scale and simplicity are not mutually exclusive. Big Data teaches you to build big data systems using an architecture designed specifically to capture and analyze web-scale data. This book presents the Lambda Architecture, a scalable, easy-to-understand approach that can be built and run by a small team. You'll explore the theory of big data systems and how to implement them in practice. In addition to discovering a general framework for processing big data, you'll learn specific technologies like Hadoop, Storm, and NoSQL databases. This book requires no previous exposure to large-scale data analysis or NoSQL tools. Familiarity with traditional databases is helpful. What's Inside Introduction to big data systems Real-time processing of web-scale data Tools like Hadoop, Cassandra, and Storm Extensions to traditional database skills About the Authors Nathan Marz is the creator of Apache Storm and the originator of the Lambda Architecture for big data systems. James Warren is an analytics architect with a background in machine learning and scientific computing. Table of Contents A new paradigm for Big Data PART 1 BATCH LAYER Data model for Big Data Data model for Big Data: Illustration Data storage on the batch layer Data storage on the batch layer: Illustration Batch layer Batch layer: Illustration An example batch layer: Architecture and algorithms An example batch layer: Implementation PART 2 SERVING LAYER Serving layer Serving layer: Illustration PART 3 SPEED LAYER Realtime views Realtime views: Illustration Queuing and stream processing Queuing and stream processing: Illustration Micro-batch stream processing Micro-batch stream processing: Illustration Lambda Architecture in depth

Apache Hive Essentials - Dayong Du 2018-06-30

This book takes you on a fantastic journey to discover the attributes of big data using Apache Hive. Key Features Grasp the skills needed to write efficient Hive queries to analyze the Big Data Discover how Hive can coexist and work with other tools within the Hadoop ecosystem Uses practical, example-oriented scenarios to cover all the newly released features of Apache Hive 2.3.3 Book Description In this book, we prepare you for your journey into big data by frstly introducing you to backgrounds in the big data domain, alongwith the process of setting up and getting familiar with your Hive working environment. Next, the book guides you through discovering and transforming the values of big data with the help of examples. It also hones your skills in using the Hive language in an effcient manner. Toward the end, the book focuses on advanced topics, such as performance, security, and extensions in Hive, which will guide you on exciting adventures on this worthwhile big data journey. By the end of the book, you will be familiar with Hive and able to work effeciently to find solutions to big data problems What you will learn Create and set up the Hive environment Discover how to use Hive's definition language to describe data Discover interesting data by joining and filtering datasets in Hive Transform data by using Hive sorting, ordering, and functions Aggregate and sample data in different ways Boost Hive query performance and enhance data security in Hive Customize Hive to your needs by using user-defined functions and integrate it with other tools Who this book is for If you are a data analyst, developer, or simply someone who wants to quickly get started with Hive to explore and analyze Big Data in Hadoop, this is the book for you. Since Hive is an SQL-like language, some previous experience with SQL will be useful to get the most out of this book.

Apache Hadoop YARN - Arun C. Murthy 2014

"Apache Hadoop is helping drive the Big Data revolution. Now, its data processing has been completely overhauled: Apache Hadoop YARN provides resource management at data center scale and easier ways to create distributed applications that process petabytes of data. And now in Apache HadoopTM YARN, two Hadoop technical leaders show you how to develop new applications and adapt existing code to fully leverage these revolutionary advances." -- From the Amazon

Introducing Data Science - Davy Cielen 2016-05-02

Summary Introducing Data Science teaches you how to accomplish the fundamental tasks that occupy data scientists. Using the Python language and common Python libraries, you'll experience firsthand the challenges of dealing with data at scale and gain a solid foundation in data science. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Many companies need developers with data science skills to work on projects ranging from social media marketing to machine learning. Discovering what you need to learn to begin a career as a data scientist can seem bewildering. This book is designed to help you get started. About the Book Introducing Data ScienceIntroducing Data Science explains vital data science concepts and teaches you how to accomplish the fundamental tasks that occupy data scientists. You'll explore data visualization, graph databases, the use of NoSQL, and the data science process. You'll use the Python language and common Python libraries as you experience firsthand the challenges of dealing with data at scale. Discover how Python allows you to gain insights from data sets so big that they need to be stored on multiple machines, or from data moving so quickly that no single machine can handle it. This book gives you hands-on experience with the most popular Python data science libraries, Scikit-learn and StatsModels. After reading this book, you'll have the solid foundation you need to start a career in data science. What's Inside Handling large data Introduction to machine learning Using Python to work with data Writing data science algorithms About the Reader This book assumes you're comfortable reading code in Python or a similar language, such as C, Ruby, or JavaScript. No prior experience with data science is required. About the Authors Davy Cielen, Arno D. B. Meysman, and Mohamed Ali are the founders and managing partners of Optimately and Maiton, where they focus on developing data science projects and solutions in various sectors. Table of Contents Data science in a big data world The data science process Machine learning Handling large data on a single computer First steps in big data Join the NoSQL movement The rise of

graph databases Text mining and text analytics Data visualization to the end user

Genetic Algorithms in Search, Optimization, and Machine Learning - David Edward Goldberg 1989

A gentle introduction to genetic algorithms. Genetic algorithms revisited: mathematical foundations. Computer implementation of a genetic algorithm. Some applications of genetic algorithms. Advanced operators and techniques in genetic search. Introduction to genetics-based machine learning. Applications of genetics-based machine learning. A look back, a glance ahead. A review of combinatorics and elementary probability. Pascal with random number generation for fortran, basic, and cobol programmers. A simple genetic algorithm (SGA) in pascal. A simple classifier system(SCS) in pascal. Partition coefficient transforms for problem-coding analysis.

The Robotic Process Automation Handbook - Tom Taulli 2020-02-28

While Robotic Process Automation (RPA) has been around for about 20 years, it has hit an inflection point because of the convergence of cloud computing, big data and AI. This book shows you how to leverage RPA effectively in your company to automate repetitive and rules-based processes, such as scheduling, inputting/transferring data, cut and paste, filling out forms, and search. Using practical aspects of implementing the technology (based on case studies and industry best practices), you'll see how companies have been able to realize substantial ROI (Return On Investment) with their implementations, such as by lessening the need for hiring or outsourcing. By understanding the core concepts of RPA, you'll also see that the technology significantly increases compliance – leading to fewer issues with regulations – and minimizes costly errors. RPA software revenues have recently soared by over 60 percent, which is the fastest ramp in the tech industry, and they are expected to exceed $1 billion by the end of 2019. It is generally seamless with legacy IT environments, making it easier for companies to pursue a strategy of digital transformation and can even be a gateway to AI. The Robotic Process Automation Handbook puts everything you need to know into one place to be a part of this wave. What You'll Learn Develop the right strategy and planDeal with resistance and fears from employeesTake an in-depth look at the leading RPA systems, including where they are most effective, the risks and the costsEvaluate an RPA system Who This Book Is For IT specialists and managers at mid-to-large companies

Apache Hive Cookbook - Hanish Bansal 2016-04-29

Easy, hands-on recipes to help you understand Hive and its integration with frameworks that are used widely in today's big data world About This Book Grasp a complete reference of different Hive topics. Get to know the latest recipes in development in Hive including CRUD operations Understand Hive internals and integration of Hive with different frameworks used in today's world. Who This Book Is For The book is intended for those who want to start in Hive or who have basic understanding of Hive framework. Prior knowledge of basic SQL command is also required What You Will Learn Learn different features and offering on the latest Hive Understand the working and structure of the Hive internals Get an insight on the latest development in Hive framework Grasp the concepts of Hive Data Model Master the key concepts like Partition, Buckets and Statistics Know how to integrate Hive with other frameworks such as Spark, Accumulo, etc In Detail Hive was developed by Facebook and later open sourced in Apache community. Hive provides SQL like interface to run queries on Big Data frameworks. Hive provides SQL like syntax also called as HiveQL that includes all SQL capabilities like analytical functions which are the need of the hour in today's Big Data world. This book provides you easy installation steps with different types of metastores supported by Hive. This book has simple and easy to learn recipes for configuring Hive clients and services. You would also learn different Hive optimizations including Partitions and Bucketing. The book also covers the source code explanation of latest Hive version. Hive Query Language is being used by other frameworks including spark. Towards the end you will cover integration of Hive with these frameworks. Style and approach Starting with the basics and covering the core concepts with the practical usage, this book is a complete guide to learn and explore Hive offerings.

Hadoop: The Definitive Guide - Tom White 2012-05-10

Ready to unlock the power of your data? With this comprehensive guide, you'll learn how to build and maintain reliable, scalable, distributed systems with Apache Hadoop. This book is ideal for programmers looking to analyze datasets of any size, and for administrators who want to set up and run Hadoop clusters. You'll find illuminating case studies that demonstrate how Hadoop is used to solve specific problems. This third edition covers recent changes to Hadoop, including material on the new MapReduce API, as well as MapReduce 2 and its more flexible execution model (YARN). Store large datasets with the Hadoop Distributed File System (HDFS) Run distributed computations with MapReduce Use Hadoop's data and I/O building blocks for compression, data integrity, serialization (including Avro), and persistence Discover common pitfalls and advanced features for writing real-world MapReduce programs Design, build, and administer a dedicated Hadoop cluster—or run Hadoop in the cloud Load data from relational databases into HDFS, using Sqoop Perform large-scale data processing with the Pig query language Analyze datasets with Hive, Hadoop's data warehousing system Take advantage of HBase for structured and semi-structured data, and ZooKeeper for building distributed systems

Between Truth and Power - Julie E. Cohen 2019

This work explores the relationships between legal institutions and political and economic transformation. It argues that as law is enlisted to help produce the profound economic and sociotechnical shifts that have accompanied the emergence of the informational economy, it is changing in fundamental ways.

Hadoop in Action - Chuck Lam 2010-11-30

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.

Big Data Analytics - David Loshin 2013-08-23

Big Data Analytics will assist managers in providing an overview of the drivers for introducing big data technology into the organization and for understanding the types of business problems best suited to big data analytics solutions, understanding the value drivers and benefits, strategic planning, developing a pilot, and eventually planning to integrate back into production within the enterprise. Guides the reader in assessing the opportunities and value proposition Overview of big data hardware and software architectures Presents a variety of technologies and how they fit into the big data ecosystem

SOA Patterns - Arnon Rotem-Gal-Oz 2012-09-11

Summary SOA Patterns provides architectural guidance through patterns and antipatterns. It shows you how to build real SOA services that feature flexibility, availability, and scalability. Through an extensive set of patterns, this book identifies the major SOA pressure points and provides reusable techniques to address them. Each pattern pairs the classic problem/solution format with a unique technology map, showing where specific solutions fit into the general pattern. About the Technology The idea of service-oriented architecture is an easy one to grasp and yet developers and enterprise architects often struggle with implementation issues. Here are some of them: How to get high availability and high performance How to know a service has failed How to create reports when data is scattered within multiple services How to make loose coupling looser How to solve authentication and authorization for service consumers How to integrate SOA and the UI About the Book SOA Patterns provides detailed, technology-neutral solutions to these challenges, and many others, using plain language. You'll understand the design patterns that promote and enforce flexibility, availability, and scalability. Each of the 26 patterns uses the classic problem/solution format and a unique technology map to show where specific solutions fit into the general

pattern. The book is written for working developers and architects building services and service-oriented solutions. Knowledge of Java or C# is helpful but not required. 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. Table of Contents PART 1 SOA PATTERNS Solving SOA pains with patterns Foundation structural patterns Patterns for performance, scalability, and availability Security and manageability patterns Message exchange patterns Service consumer patterns Service integration patterns PART 2 SOA IN THE REAL WORLD Service antipatterns Putting it all together—a case study SOA vs. the world

From Mathematics to Generic Programming - Alexander A. Stepanov 2014-11-13

In this substantive yet accessible book, pioneering software designer Alexander Stepanov and his colleague Daniel Rose illuminate the principles of generic programming and the mathematical concept of abstraction on which it is based, helping you write code that is both simpler and more powerful. If you’re a reasonably proficient programmer who can think logically, you have all the background you’ll need. Stepanov and Rose introduce the relevant abstract algebra and number theory with exceptional clarity. They carefully explain the problems mathematicians first needed to solve, and then show how these mathematical solutions translate to generic programming and the creation of more effective and elegant code. To demonstrate the crucial role these mathematical principles play in many modern applications, the authors show how to use these results and generalized algorithms to implement a real-world public-key cryptosystem. As you read this book, you’ll master the thought processes necessary for effective programming and learn how to generalize narrowly conceived algorithms to widen their usefulness without losing efficiency. You’ll also gain deep insight into the value of mathematics to programming—insight that will prove invaluable no matter what programming languages and paradigms you use. You will learn about How to generalize a four thousand-year-old algorithm, demonstrating indispensable lessons about clarity and efficiency Ancient paradoxes, beautiful theorems, and the productive tension between continuous and discrete A simple algorithm for finding greatest common divisor (GCD) and modern abstractions that build on it Powerful mathematical approaches to abstraction How abstract algebra provides the idea at the heart of generic programming Axioms, proofs, theories, and models: using mathematical techniques to organize knowledge about your algorithms and data structures Surprising subtleties of simple programming tasks and what you can learn from them How practical implementations can exploit theoretical knowledge

Linux Clustering - Charles Bookman 2003

"Linux Clustering" is the premier resource for system administrators wishing to implement clustering solutions on the many types of Linux systems. It guides Linux Administrators through difficult tasks while offering helpful tips and tricks.

Hadoop Operations - Eric Sammer 2012-09-26

If you’ve been asked to maintain large and complex Hadoop clusters, this book is a must. Demand for operations-specific material has skyrocketed now that Hadoop is becoming the de facto standard for truly large-scale data processing in the data center. Eric Sammer, Principal Solution Architect at Cloudera, shows you the particulars of running Hadoop in production, from planning, installing, and configuring the system to providing ongoing maintenance. Rather than run through all possible scenarios, this pragmatic operations guide calls out what works, as demonstrated in critical deployments. Get a high-level overview of HDFS and MapReduce: why they exist and how they work Plan a Hadoop deployment, from hardware and OS selection to network requirements Learn setup and configuration details with a list of critical properties Manage resources by sharing a cluster across multiple groups Get a runbook of the most common cluster maintenance tasks Monitor Hadoop clusters—and learn troubleshooting with the help of real-world war stories Use basic tools and techniques to handle backup and catastrophic failure

Spark in Action - Marko Bonaci 2016-11-03

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

Hadoop in Practice - Alex Holmes 2014-09-29

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

Hadoop in Practice - Alex Holmes 2015

Data Analytics with Hadoop - Benjamin Bengfort 2016-06

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

Re-imagining Diffusion and Adoption of Information Technology and Systems: A Continuing Conversation - Sujeet K. Sharma 2020-12-16

This two-volume set of IFIP AICT 617 and 618 constitutes the refereed proceedings of the IFIP WG 8.6 International Working Conference "Re-imagining Diffusion and Adoption of Information Technology and Systems: A Continuing Conversation" on Transfer and Diffusion of IT, TDIT 2020, held in Tiruchirappalli, India, in December 2020. The 86 revised full papers and 36 short papers presented were carefully reviewed and selected from 224 submissions. The papers focus on the re-imagination of diffusion and adoption of emerging technologies. They are organized in the following parts: Part I: artificial intelligence and autonomous systems; big data and analytics; blockchain; diffusion and adoption technology; emerging technologies in e-Governance; emerging technologies in consumer decision making and choice; fin-tech applications; healthcare information technology; and Internet of Things Part II: diffusion of information technology and disaster management; adoption of mobile and platform-based applications; smart cities and digital government; social media; and diffusion of information technology and systems

Mastering Elasticsearch - Second Edition - Rafał Kuć 2015-02-27

This book is for Elasticsearch users who want to extend their knowledge and develop new skills. Prior knowledge of the Query DSL and data indexing is expected.

Hadoop Beginner's Guide - Garry Turkington 2013-02-22

Data is arriving faster than you can process it and the overall volumes keep growing at a rate that keeps you awake at night. Hadoop can help you tame the data beast. Effective use of Hadoop however requires a mixture of programming, design, and system administration skills. "Hadoop Beginner's Guide" removes the mystery from Hadoop, presenting Hadoop and related technologies with a focus on building working systems and getting the job done, using cloud services to do so when it makes sense. From basic concepts and initial setup through developing applications and keeping the system running as the data grows, the book gives the understanding needed to effectively use Hadoop to solve real world problems. Starting with the basics of installing and configuring Hadoop, the book explains how to develop applications, maintain the system, and how to use additional products to integrate with other systems. While learning different ways to develop applications to run on Hadoop the book also covers tools such as Hive, Sqoop, and Flume that show how Hadoop can be integrated with relational databases and log collection. In addition to examples on Hadoop clusters on Ubuntu uses of cloud services such as Amazon, EC2 and Elastic MapReduce are covered.

Hadoop Application Architectures - Mark Grover 2015-06-30

Get expert guidance on architecting end-to-end data management solutions with Apache Hadoop. While many sources explain how to use various components in the Hadoop ecosystem, this practical book takes you through architectural considerations necessary to tie those components together into a complete tailored application, based on your particular use case. To reinforce those lessons, the book's second section provides detailed examples of architectures used in some of the most commonly found Hadoop applications. Whether you're designing a new Hadoop application, or planning to integrate Hadoop into your existing data infrastructure, Hadoop Application Architectures will skillfully guide you through the process. This book covers: Factors to consider when using Hadoop to store and model data Best practices for moving data in and out of the system Data processing frameworks, including MapReduce, Spark, and Hive Common Hadoop processing patterns, such as removing duplicate records and using windowing analytics Giraph, GraphX, and other tools for large graph processing on Hadoop Using workflow orchestration and scheduling tools such as Apache Oozie Near-real-time stream processing with Apache Storm, Apache Spark Streaming, and Apache Flume Architecture examples for clickstream analysis, fraud detection, and data warehousing

MapReduce Design Patterns - Donald Miner 2012-11-21

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

Practical Hadoop Ecosystem - Deepak Vohra 2016-09-30

Learn how to use the Apache Hadoop projects, including MapReduce, HDFS, Apache Hive, Apache HBase, Apache Kafka, Apache Mahout, and Apache Solr. From setting up the environment to running sample applications each chapter in this book is a practical tutorial on using an Apache Hadoop ecosystem project. While several books on Apache Hadoop are available, most are based on the main projects, MapReduce and HDFS, and none discusses the other Apache Hadoop ecosystem projects and how they all work together as a cohesive big data development platform. What You Will Learn: Set up the environment in Linux for Hadoop projects using Cloudera Hadoop Distribution CDH 5 Run a MapReduce job Store data with Apache Hive, and Apache HBase Index data in HDFS with Apache Solr Develop a Kafka messaging system Stream Logs to HDFS with Apache Flume Transfer data from MySQL database to Hive, HDFS, and HBase with Sqoop Create a Hive table over Apache Solr Develop a Mahout User Recommender System Who This Book Is For: Apache Hadoop developers. Pre-requisite knowledge of Linux and some knowledge of Hadoop is required.