20 Essential Hadoop Tools For Crunching Big Data

Yeah, reviewing a book **20 essential hadoop tools for crunching big data** could accumulate your close associates listings. This is just one of the solutions for you to be successful. As understood, ability does not suggest that you have fantastic points.

Comprehending as capably as union even more than additional will give each success. bordering to, the message as competently as sharpness of this 20 essential hadoop tools for crunching big data can be taken as skillfully as picked to act.

Expert Hadoop 2 Administration Sam R. Alapati 2016-11-29 This is the eBook of the printed book and may not include any media, website access codes, or print supplements that may come packaged with the bound book. The Comprehensive, Up-to-Date Apache Hadoop Administration Handbook and Reference "Sam Alapati has worked with production Hadoop clusters for six years. His unique depth of experience has enabled him to write the go-to resource for all administrators looking to spec, size, expand, and secure production Hadoop clusters of any size." —Paul Dix, Series Editor In Expert Hadoop® Administration, leading Hadoop administrator Sam R. Alapati brings together authoritative knowledge for creating, configuring, securing, managing, and optimizing production Hadoop clusters in any environment. Drawing on his experience with large-scale Hadoop administration, Alapati integrates action-oriented advice with carefully researched explanations of both problems and solutions. He covers an unmatched range of topics and offers an unparalleled collection of realistic examples. Alapati demystifies complex Hadoop environments, helping you understand exactly what happens behind the scenes when you administer your cluster. You'll gain unprecedented insight as you walk through building clusters from scratch and configuring high availability, performance, security, encryption, and other key attributes. The high-value administration skills you learn here will be indispensable no matter what Hadoop distribution you use or what Hadoop applications you run. Understand Hadoop's architecture from an administrator's standpoint Create simple and fully distributed clusters Run MapReduce and Spark applications in a Hadoop cluster Manage and protect Hadoop data and high availability Work with HDFS commands, file permissions, and storage management Move data, and use YARN to allocate resources and schedule jobs Manage job workflows with Oozie and Hue Secure, monitor, log, and optimize Hadoop Benchmark and troubleshoot Hadoop

Python for Data Analysis Wes McKinney 2017-09-25 Get complete instructions for manipulating, processing, cleaning, and crunching datasets in Python. Updated for Python 3.6, the second edition of this hands-on guide is packed with practical case studies that show you how to solve a broad set of data analysis problems effectively. You'll learn the latest versions of pandas, NumPy, IPython, and Jupyter in the process. Written by Wes McKinney, the creator of the Python pandas project, this book is a practical, modern introduction to data science tools in Python. It's ideal for analysts new to Python and for Python programmers new to data science and scientific computing. Data files and related material are available on GitHub. Use the IPython shell and Jupyter notebook for exploratory computing Learn basic and advanced features in NumPy (Numerical Python) Get started with data analysis tools in the pandas library Use flexible tools to load, clean, transform, merge, and reshape data Create informative visualizations with matplotlib Apply the pandas groupby facility to slice, dice, and summarize datasets Analyze and manipulate regular and irregular time series data Learn how to solve real-world data analysis problems with thorough, detailed examples

Designing Data-Intensive Applications Martin Kleppmann 2017-03-16 Data is at the center of many challenges in system design today. Difficult issues need to be figured out, such as scalability, consistency, reliability, efficiency, and maintainability. In addition, we have an overwhelming variety of tools, including relational databases, NoSQL datastores, stream or batch processors, and message brokers. What are the right choices for your application? How do you make sense of all these buzzwords? In this practical and comprehensive guide, author Martin Kleppmann helps you navigate this diverse landscape by examining the pros and cons of various technologies for processing and storing data. Software keeps changing, but the fundamental principles remain the same. With this book, software engineers and architects will learn how to apply those ideas in practice, and how to make full use of data in modern applications. Peer under the hood of the systems you already use, and learn how to use and operate them more effectively Make informed decisions by identifying the strengths and weaknesses of different tools Navigate the trade-offs around consistency, scalability, fault tolerance, and complexity Understand the distributed systems research upon which modern databases are built Peek behind the scenes of major online services, and learn from their architectures

MKTG Charles W. Lamb 2017-01-27 4LTR Press solutions give students the option to choose the format that best suits their learning preferences. This option is perfect for those students who focus on the textbook as their main course resource. Important Notice: Media content referenced within the product description or the product text may not be available in the ebook version.

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.

Big Data Analytics with Java Rajat Mehta 2017-07-31 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.

HBase Lars George 2011-09-05 If your organization is looking for a storage solution to accommodate a virtually endless amount of data, this book will show you how Apache HBase can fulfill your needs. As the open source implementation of Google's BigTable architecture, HBase scales to billions of rows and millions of columns, while ensuring that write and read performance remain constant.HBase: The Definitive Guideprovides the details you require, whether you simply want to evaluate this high-performance, non-relational database, or put it into practice right away. HBase's adoption rate is beginning to climb, and several IT executives are asking pointed questions about this high-capacity database. This is the only book available to give you meaningful answers. Learn how to distribute large datasets across an inexpensive cluster of commodity servers Develop HBase clients in many programming languages, including Java, Python, and Ruby Get details on HBase's primary storage system, HDFS—Hadoop's distributed and replicated filesystem Learn how HBase's native interface to Hadoop's MapReduce framework enables easy development and execution of batch jobs that can scan entire tables Discover the integration between HBase and other facets of the Apache Hadoop project

Big Data Concepts, Theories, and Applications Shui Yu 2016-03-03 This book covers three major parts of Big Data: concepts, theories and applications. Written by world-renowned leaders in Big Data, this book explores the problems, possible solutions and directions for Big Data in research and practice. It also focuses on high level concepts such as definitions of Big Data from different angles; surveys in research and applications; and existing tools, mechanisms, and systems in practice. Each chapter is independent from the other chapters, allowing users to read any chapter directly. After examining the practical side of Big Data, this book presents theoretical perspectives. The theoretical research ranges from Big Data representation, modeling and topology to distribution and dimension reducing. Chapters also investigate the many disciplines that involve Big Data, such as statistics, data mining, machine learning, networking, algorithms, security and differential geometry. The last section of this book introduces Big Data applications from different communities, such as business, engineering and science. Big Data Concepts, Theories and Applications is designed as a reference for researchers and advanced level students in computer science, electrical engineering and mathematics. Practitioners who focus on information systems, big data, data mining, business analysis and other related fields will also find this material valuable.

Hadoop: The Definitive Guide Tom White 2010-09-24 Discover how Apache Hadoop can unleash the power of your data. This comprehensive resource shows you how to build and maintain reliable, scalable, distributed systems with the Hadoop framework -- an open source implementation of MapReduce, the algorithm on which Google built its empire. Programmers will find details for analyzing datasets of any size, and administrators will learn how to set up and run Hadoop clusters. This revised edition covers recent changes to Hadoop, including new features such as Hive, Sqoop, and Avro. It also provides illuminating case studies that illustrate how Hadoop is used to solve specific problems. Looking to get the most out of your data? This is your book. Use the Hadoop Distributed File System (HDFS) for storing large datasets, then run distributed computations over those datasets with MapReduce Become familiar with Hadoop's data and I/O building blocks for compression, data integrity, serialization, 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 Use Pig, a high-level query language for large-scale data processing Analyze datasets with Hive, Hadoop's data warehousing system Take advantage of HBase, Hadoop's database for structured and semi-structured data Learn ZooKeeper, a toolkit of coordination primitives for building distributed systems "Now you have the opportunity to learn about Hadoop from a master -- not only of the technology, but also of common sense and plain talk." --Doug Cutting, Cloudera

Hadoop Operations and Cluster Management Cookbook Shumin Guo 2013 Solve specific problems using individual self-contained code recipes, or work through the book to develop your capabilities. This book is packed with easy-to-follow code and commands used for illustration, which makes your learning curve easy and quick.If you are a Hadoop cluster system administrator with Unix/Linux system management experience and you are looking to get a good grounding in how to set up and manage a Hadoop cluster, then this book is for you. It's assumed that you will have some

experience in Unix/Linux command line already, as well as being familiar with network communication basics.

Mastering Hadoop 3 Chanchal Singh 2019-02-28 A comprehensive guide to mastering the most advanced Hadoop 3 concepts Key Features Get to grips with the newly introduced features and capabilities of Hadoop 3 Crunch and process data using MapReduce, YARN, and a host of tools within the Hadoop ecosystem Sharpen your Hadoop skills with real-world case studies and code Book Description Apache Hadoop is one of the most popular big data solutions for distributed storage and for processing large chunks of data. With Hadoop 3, Apache promises to provide a high-performance, more fault-tolerant, and highly efficient big data processing platform, with a focus on improved scalability and increased efficiency. With this guide, you'll understand advanced concepts of the Hadoop ecosystem tool. You'll learn how Hadoop works internally, study advanced concepts of different ecosystem tools, discover solutions to real-world use cases, and understand how to secure your cluster. It will then walk you through HDFS, YARN, MapReduce, and Hadoop 3 concepts. You'll be able to address common challenges like using Kafka efficiently, designing low latency, reliable message delivery Kafka systems, and handling high data volumes. As you advance, you'll discover how to address major challenges when building an enterprise-grade messaging system, and how to use different stream processing systems along with Kafka to fulfil your enterprise goals. By the end of this book, you'll have a complete understanding of how components in the Hadoop ecosystem are effectively integrated to implement a fast and reliable data pipeline, and you'll be equipped to tackle a range of real-world problems in data pipelines. What you will learn Gain an in-depth understanding of distributed computing using Hadoop 3 Develop enterprise-grade applications using Apache Spark, Flink, and more Build scalable and high-performance Hadoop data pipelines with security, monitoring, and data governance Explore batch data processing patterns and how to model data in Hadoop Master best practices for enterprises using, or planning to use, Hadoop 3 as a data platform Understand security aspects of Hadoop, including authorization and authentication Who this book is for If you want to become a big data professional by mastering the advanced concepts of Hadoop, this book is for you. You'll also find this book useful if you're a Hadoop professional looking to strengthen your knowledge of the Hadoop ecosystem. Fundamental knowledge of the Java programming language and basics of Hadoop is necessary to get started with this book.

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

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

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.

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

Social Media Listening and Monitoring for Business Applications Rao, N. Raghavendra 2016-09-21 Social Media has transformed the ways in which individuals keep in touch with family and friends. Likewise, businesses have identified the profound opportunities present for customer engagement and understanding through the massive data available on social media channels, in addition to the customer reach of such sites. Social Media Listening and Monitoring for Business Applications explores research-based solutions for businesses of all types interested in an understanding of emerging concepts and technologies for engaging customers online. Providing insight into the currently available social media tools and practices for various business applications, this publication is an essential resource for business professionals, graduate-level students, technology developers, and researchers.

Big Data Rajkumar Buyya 2016-06-07 Big Data: Principles and Paradigms captures the state-of-the-art research on the architectural aspects, technologies, and applications of Big Data. The book identifies potential future directions and technologies that facilitate insight into numerous scientific, business, and consumer applications. To help realize Big Data's full potential, the book addresses numerous challenges, offering the conceptual and technological solutions for tackling them. These challenges include life-cycle data management, large-scale storage, flexible processing infrastructure, data modeling, scalable machine learning, data analysis algorithms, sampling techniques, and privacy and ethical issues. Covers computational platforms supporting Big Data applications Addresses key principles underlying Big Data computing Examines key developments supporting next generation Big Data platforms Explores the challenges in Big Data computing and ways to overcome them Contains expert contributors from both academia and industry

Practical Data Science with Hadoop and Spark Ofer Mendelevitch 2016-12-08 The Complete Guide to Data Science with Hadoop—For Technical Professionals, Businesspeople, and Students Demand is soaring for professionals who can solve real data science problems with Hadoop and Spark. Practical Data Science with Hadoop® and Spark is your complete guide to doing just that. Drawing on immense experience with Hadoop and big data, three leading experts bring together everything you need: high-level concepts, deep-dive techniques, real-world use cases, practical applications, and hands-on tutorials. The authors introduce the essentials of data science and the modern Hadoop ecosystem, explaining how Hadoop

and Spark have evolved into an effective platform for solving data science problems at scale. In addition to comprehensive application coverage, the authors also provide useful guidance on the important steps of data ingestion, data munging, and visualization. Once the groundwork is in place, the authors focus on specific applications, including machine learning, predictive modeling for sentiment analysis, clustering for document analysis, anomaly detection, and natural language processing (NLP). This guide provides a strong technical foundation for those who want to do practical data science, and also presents business-driven guidance on how to apply Hadoop and Spark to optimize ROI of data science initiatives. Learn What data science is, how it has evolved, and how to plan a data science career How data volume, variety, and velocity shape data science use cases Hadoop and its ecosystem, including HDFS, MapReduce, YARN, and Spark Data importation with Hive and Spark Data quality, preprocessing, preparation, and modeling Visualization: surfacing insights from huge data sets Machine learning: classification, regression, clustering, and anomaly detection Algorithms and Hadoop tools for predictive modeling Cluster analysis and similarity functions Large-scale anomaly detection NLP: applying data science to human language

Big Data Analytics Venkat Ankam 2016-09-28 A handy reference guide for data analysts and data scientists to help to obtain value from big data analytics using Spark on Hadoop clusters About This Book This book is based on the latest 2.0 version of Apache Spark and 2.7 version of Hadoop integrated with most commonly used tools. Learn all Spark stack components including latest topics such as DataFrames, DataSets, GraphFrames, Structured Streaming, DataFrame based ML Pipelines and SparkR. Integrations with frameworks such as HDFS, YARN and tools such as Jupyter, Zeppelin, NiFi, Mahout, HBase Spark Connector, GraphFrames, H2O and Hivemall. Who This Book Is For Though this book is primarily aimed at data analysts and data scientists, it will also help architects, programmers, and practitioners. Knowledge of either Spark or Hadoop would be beneficial. It is assumed that you have basic programming background in Scala, Python, SQL, or R programming with basic Linux experience. Working experience within big data environments is not mandatory. What You Will Learn Find out and implement the tools and techniques of big data analytics using Spark on Hadoop clusters with wide variety of tools used with Spark and Hadoop Understand all the Hadoop and Spark ecosystem components Get to know all the Spark components: Spark Core, Spark SQL, DataFrames, DataSets, Conventional and Structured Streaming, MLLib, ML Pipelines and Graphx See batch and real-time data analytics using Spark Core, Spark SQL, and Conventional and Structured Streaming Get to grips with data science and machine learning using MLLib, ML Pipelines, H2O, Hivemall, Graphx, SparkR and Hivemall. In Detail Big Data Analytics book aims at providing the fundamentals of Apache Spark and Hadoop. All Spark components – Spark Core, Spark SQL, DataFrames, Data sets, Conventional Streaming, Structured Streaming, MLlib, Graphx and Hadoop core components – HDFS, MapReduce and Yarn are explored in greater depth with implementation examples on Spark + Hadoop clusters. It is moving away from MapReduce to Spark. So, advantages of Spark over MapReduce are explained at great depth to reap benefits of in-memory speeds. DataFrames API, Data Sources API and new Data set API are explained for building Big Data analytical applications. Real-time data analytics using Spark Streaming with Apache Kafka and HBase is covered to help building streaming applications. New Structured streaming concept is explained with an IOT (Internet of Things) use case. Machine learning techniques are covered using MLLib, ML Pipelines and SparkR

and Graph Analytics are covered with GraphX and GraphFrames components of Spark. Readers will also get an opportunity to get started with web based notebooks such as Jupyter, Apache Zeppelin and data flow tool Apache NiFi to analyze and visualize data. Style and approach This step-by-step pragmatic guide will make life easy no matter what your level of experience. You will deep dive into Apache Spark on Hadoop clusters through ample exciting real-life examples. Practical tutorial explains data science in simple terms to help programmers and data analysts get started with Data Science

Digital Journalism, Drones, and Automation Cate Dowd 2020-01-27 The lure of big data and analytics has produced new partnerships between news media and social media and consequently a fragmentation of digital journalism. The era is coupled with the rise in fake news and controversial data sharing. However, creative mobile reporting and civilian drones set new standards for journalist during the European asylum seeker crisis. Yet the focus on data and remote cloud servers continues to dominate online news and journalism, alongside new semantic models for data personalization. News tags that define concepts within a news story to assist search, are now monetized abstractions in accelerated data processing that enables automation and feeds advertising. Can journalism compete with this by defining its own concepts with ethical values named and embedded in algorithms? Can machines make sense of the world in the same way as a traditional journalist? In this book, Cate Dowd analyzes the tasks and ethics of journalists and questions how intelligent machines could simulate ethical human behaviors to better understand the dizzy post-human world of online data. Looking to digital journalism and multi-platform news media, from studios and integrated media systems to mobile reporting in the field, Dowd assesses how data and digital technology has impacted on journalism over the past decade. Dowd's research is informed by in-depth participation with investigative journalists, including images drawn and annotated by industry experts to present key journalism concepts, priorities, and values. Chapters explore approaches for the elicitation of vocabulary for journalism and design methods to embed values and ethics into algorithms for the era of automation and big data. Digital Journalism, Drones, and Automation provides insights into the lasting values of journalism processes and equips readers interested in entering or understanding online data and news media with much needed context and wisdom.

HADOOP TOM. WHITE 2015

Next Generation Databases Guy Harrison 2015-12-30 "It's not easy to find such a generous book on big data and databases. Fortunately, this book is the one." Feng Yu. Computing Reviews. June 28, 2016. This is a book for enterprise architects, database administrators, and developers who need to understand the latest developments in database technologies. It is the book to help you choose the correct database technology at a time when concepts such as Big Data, NoSQL and NewSQL are making what used to be an easy choice into a complex decision with significant implications. The relational database (RDBMS) model completely dominated database technology for over 20 years. Today this "one size fits all" stability has been disrupted by a relatively recent explosion of new database technologies. These paradigm-busting technologies are powering the "Big Data" and "NoSQL" revolutions, as well as forcing fundamental changes in databases across the board. Deciding to use a relational database was once truly a no-brainer, and the various commercial relational databases competed on price, performance, reliability, and ease of use rather than on fundamental architectures. Today we are faced with choices between radically different database technologies. Choosing

the right database today is a complex undertaking, with serious economic and technological consequences. Next Generation Databases demystifies today's new database technologies. The book describes what each technology was designed to solve. It shows how each technology can be used to solve real word application and business problems. Most importantly, this book highlights the architectural differences between technologies that are the critical factors to consider when choosing a database platform for new and upcoming projects. Introduces the new technologies that have revolutionized the database landscape Describes how each technology can be used to solve specific application or business challenges Reviews the most popular new wave databases and how they use these new database technologies

Hadoop: The Definitive Guide Tom White 2015-03-25 Get ready to unlock the power of your data. With the fourth edition of 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. Using Hadoop 2 exclusively, author Tom White presents new chapters on YARN and several Hadoop-related projects such as Parquet, Flume, Crunch, and Spark. You'll learn about recent changes to Hadoop, and explore new case studies on Hadoop's role in healthcare systems and genomics data processing. Learn fundamental components such as MapReduce, HDFS, and YARN Explore MapReduce in depth, including steps for developing applications with it Set up and maintain a Hadoop cluster running HDFS and MapReduce on YARN Learn two data formats: Avro for data serialization and Parquet for nested data Use data ingestion tools such as Flume (for streaming data) and Sqoop (for bulk data transfer) Understand how high-level data processing tools like Pig, Hive, Crunch, and Spark work with Hadoop Learn the HBase distributed database and the ZooKeeper distributed configuration service

Big Data Viktor Mayer-Schonberger 2013-03-14 New and expanded edition. An International Bestseller - Over One Million Copies Sold! Shortlisted for the Financial Times/Goldman Sachs Business Book of the Year Award. Since Aristotle, we have fought to understand the causes behind everything. But this ideology is fading. In the age of big data, we can crunch an incomprehensible amount of information, providing us with invaluable insights about the what rather than the why. We're just starting to reap the benefits: tracking vital signs to foresee deadly infections, predicting building fires, anticipating the best moment to buy a plane ticket, seeing inflation in real time and monitoring social media in order to identify trends. But there is a dark side to big data. Will it be machines, rather than people, that make the decisions? How do you regulate an algorithm? What will happen to privacy? Will individuals be punished for acts they have yet to commit? In this groundbreaking and fascinating book, two of the world's most-respected data experts reveal the reality of a big data world and outline clear and actionable steps that will equip the reader with the tools needed for this next phase of human evolution.

Big Data at Work Thomas Davenport 2014-02-04 Go ahead, be skeptical about big data. The author was—at first. When the term "big data" first came on the scene, bestselling author Tom Davenport (Competing on Analytics, Analytics at Work) thought it was just another example of technology hype. But his research in the years that followed changed his mind. Now, in clear, conversational language, Davenport explains what big data means—and why everyone in business needs to know about it. Big Data at Work covers all the bases: what big data means from a technical, consumer, and management perspective; what its opportunities and costs

are; where it can have real business impact; and which aspects of this hot topic have been oversold. This book will help you understand: • Why big data is important to you and your organization • What technology you need to manage it • How big data could change your job, your company, and your industry • How to hire, rent, or develop the kinds of people who make big data work • The key success factors in implementing any big data project • How big data is leading to a new approach to managing analytics With dozens of company examples, including UPS, GE, Amazon, United Healthcare, Citigroup, and many others, this book will help you seize all opportunities—from improving decisions, products, and services to strengthening customer relationships. It will show you how to put big data to work in your own organization so that you too can harness the power of this ever-evolving new resource.

Big Data with Hadoop MapReduce Rathinaraja Jeyaraj 2020-05-01 The authors provide an understanding of big data and MapReduce by clearly presenting the basic terminologies and concepts. They have employed over 100 illustrations and many worked-out examples to convey the concepts and methods used in big data, the inner workings of MapReduce, and single node/multi-node installation on physical/virtual machines. This book covers almost all the necessary information on Hadoop MapReduce for most online certification exams. Upon completing this book, readers will find it easy to understand other big data processing tools such as Spark, Storm, etc. Ultimately, readers will be able to: • understand what big data is and the factors that are involved • understand the inner workings of MapReduce, which is essential for certification exams • learn the features and weaknesses of MapReduce • set up Hadoop clusters with 100s of physical/virtual machines • create a virtual machine in AWS • write MapReduce with Eclipse in a simple way • understand other big data processing tools and their applications

Getting Started with Kudu Jean-Marc Spaggiari 2018-07-09 Fast data ingestion, serving, and analytics in the Hadoop ecosystem have forced developers and architects to choose solutions using the least common denominator—either fast analytics at the cost of slow data ingestion or fast data ingestion at the cost of slow analytics. There is an answer to this problem. With the Apache Kudu column-oriented data store, you can easily perform fast analytics on fast data. This practical guide shows you how. Begun as an internal project at Cloudera, Kudu is an open source solution compatible with many data processing frameworks in the Hadoop environment. In this book, current and former solutions professionals from Cloudera provide use cases, examples, best practices, and sample code to help you get up to speed with Kudu. Explore Kudu's high-level design, including how it spreads data across servers Fully administer a Kudu cluster, enable security, and add or remove nodes Learn Kudu's client-side APIs, including how to integrate Apache Impala, Spark, and other frameworks for data manipulation Examine Kudu's schema design, including basic concepts and primitives necessary to make your project successful Explore case studies for using Kudu for real-time IoT analytics, predictive modeling, and in combination with another storage engine

Big Data Analytics with R and Hadoop Vignesh Prajapati 2013-11-25 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.

Big Data For Dummies Judith S. Hurwitz 2013-04-02 Find the right big data solution for your business ororganization Big data management is one of the major challenges facingbusiness, industry, and not-for-profit organizations. Data setssuch as customer transactions for a mega-retailer, weather patternsmonitored by meteorologists, or social network activity can quicklyoutpace the capacity of traditional data management tools. If youneed to develop or manage big data solutions, you'll appreciate howthese four experts define, explain, and guide you through this newand often confusing concept. You'll learn what it is, why itmatters, and how to choose and implement solutions that work. Effectively managing big data is an issue of growing importanceto businesses, not-for-profit organizations, government, and ITprofessionals Authors are experts in information management, big data, and avariety of solutions Explains big data in detail and discusses how to select andimplement a solution, security concerns to consider, data storageand presentation issues, analytics, and much more Provides essential information in a no-nonsense,easy-to-understand style that is empowering Big Data For Dummies cuts through the confusion and helpsyou take charge of big data solutions for your organization.

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

Math for Programmers Paul Orland 2021-01-12 In Math for Programmers you'll explore important mathematical concepts through hands-on coding. Filled with graphics and more than 300 exercises and mini-projects, this book unlocks the door to interesting—and lucrative!—careers in some of today's hottest fields. As you

tackle the basics of linear algebra, calculus, and machine learning, you'll master the key Python libraries used to turn them into real-world software applications. Summary To score a job in data science, machine learning, computer graphics, and cryptography, you need to bring strong math skills to the party. Math for Programmers teaches the math you need for these hot careers, concentrating on what you need to know as a developer. Filled with lots of helpful graphics and more than 200 exercises and mini-projects, this book unlocks the door to interesting—and lucrative!—careers in some of today's hottest programming fields. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Skip the mathematical jargon: This one-of-a-kind book uses Python to teach the math you need to build games, simulations, 3D graphics, and machine learning algorithms. Discover how algebra and calculus come alive when you see them in code! About the book In Math for Programmers you'll explore important mathematical concepts through hands-on coding. Filled with graphics and more than 300 exercises and mini-projects, this book unlocks the door to interesting—and lucrative!—careers in some of today's hottest fields. As you tackle the basics of linear algebra, calculus, and machine learning, you'll master the key Python libraries used to turn them into real-world software applications. What's inside Vector geometry for computer graphics Matrices and linear transformations Core concepts from calculus Simulation and optimization Image and audio processing Machine learning algorithms for regression and classification About the reader For programmers with basic skills in algebra. About the author Paul Orland is a programmer, software entrepreneur, and math enthusiast. He is co-founder of Tachyus, a start-up building predictive analytics software for the energy industry. You can find him online at www.paulor.land. Table of Contents 1 Learning math with code PART I - VECTORS AND GRAPHICS 2 Drawing with 2D vectors 3 Ascending to the 3D world 4 Transforming vectors and graphics 5 Computing transformations with matrices 6 Generalizing to higher dimensions 7 Solving systems of linear equations PART 2 - CALCULUS AND PHYSICAL SIMULATION 8 Understanding rates of change 9 Simulating moving objects 10 Working with symbolic expressions 11 Simulating force fields 12 Optimizing a physical system 13 Analyzing sound waves with a Fourier series PART 3 - MACHINE LEARNING APPLICATIONS 14 Fitting functions to data 15 Classifying data with logistic regression 16 Training neural networks

Big Data Analytics with R Simon Walkowiak 2016-07-29 Utilize R to uncover hidden patterns in your Big Data About This Book Perform computational analyses on Big Data to generate meaningful results Get a practical knowledge of R programming language while working on Big Data platforms like Hadoop, Spark, H2O and SQL/NoSQL databases, Explore fast, streaming, and scalable data analysis with the most cutting-edge technologies in the market Who This Book Is For This book is intended for Data Analysts, Scientists, Data Engineers, Statisticians, Researchers, who want to integrate R with their current or future Big Data workflows. It is assumed that readers have some experience in data analysis and understanding of data management and algorithmic processing of large quantities of data, however they may lack specific skills related to R. What You Will Learn Learn about current state of Big Data processing using R programming language and its powerful statistical capabilities Deploy Big Data analytics platforms with selected Big Data tools supported by R in a cost-effective and time-saving manner Apply the R language to real-world Big Data problems on a multi-node Hadoop cluster, e.g. electricity consumption across various socio-demographic indicators and bike share scheme usage Explore the compatibility of R with Hadoop, Spark, SQL and NoSQL

databases, and H2O platform In Detail Big Data analytics is the process of examining large and complex data sets that often exceed the computational capabilities. R is a leading programming language of data science, consisting of powerful functions to tackle all problems related to Big Data processing. The book will begin with a brief introduction to the Big Data world and its current industry standards. With introduction to the R language and presenting its development, structure, applications in real world, and its shortcomings. Book will progress towards revision of major R functions for data management and transformations. Readers will be introduce to Cloud based Big Data solutions (e.g. Amazon EC2 instances and Amazon RDS, Microsoft Azure and its HDInsight clusters) and also provide guidance on R connectivity with relational and non-relational databases such as MongoDB and HBase etc. It will further expand to include Big Data tools such as Apache Hadoop ecosystem, HDFS and MapReduce frameworks. Also other R compatible tools such as Apache Spark, its machine learning library Spark MLlib, as well as H2O. Style and approach This book will serve as a practical guide to tackling Big Data problems using R programming language and its statistical environment. Each section of the book will present you with concise and easy-to-follow steps on how to process, transform and analyse large data sets.
Machine Learning Jason Bell 2020-02-17 Dig deep into the data with a hands-on guide to machine learning with updated examples and more! Machine Learning: Hands-On for Developers and Technical Professionals provides hands-on instruction and fully-coded working examples for the most common machine learning techniques used by developers and technical professionals. The book contains a breakdown of each ML variant, explaining how it works and how it is used within certain industries, allowing readers to incorporate the presented techniques into their own work as they follow along. A core tenant of machine learning is a strong focus on data preparation, and a full exploration of the various types of learning algorithms illustrates how the proper tools can help any developer extract information and insights from existing data. The book includes a full complement of Instructor's Materials to facilitate use in the classroom, making this resource useful for students and as a professional reference. At its core, machine learning is a mathematical, algorithm-based technology that forms the basis of historical data mining and modern big data science. Scientific analysis of big data requires a working knowledge of machine learning, which forms predictions based on known properties learned from training data. Machine Learning is an accessible, comprehensive guide for the non-mathematician, providing clear guidance that allows readers to: Learn the languages of machine learning including Hadoop, Mahout, and Weka Understand decision trees, Bayesian networks, and artificial neural networks Implement Association Rule, Real Time, and Batch learning Develop a strategic plan for safe, effective, and efficient machine learning By learning to construct a system that can learn from data, readers can increase their utility across industries. Machine learning sits at the core of deep dive data analysis and visualization, which is increasingly in demand as companies discover the goldmine hiding in their existing data. For the tech professional involved in data science, Machine Learning: Hands-On for Developers and Technical Professionals provides the skills and techniques required to dig deeper.
Hadoop 2 Quick-Start Guide Douglas Eadline 2015-10-28 Get Started Fast with Apache Hadoop® 2, YARN, and Today's Hadoop Ecosystem With Hadoop 2.x and YARN, Hadoop moves beyond MapReduce to become practical for virtually any type of data processing. Hadoop 2.x and the Data Lake concept represent a radical shift away from conventional approaches to data usage and storage. Hadoop 2.x installations

offer unmatched scalability and breakthrough extensibility that supports new and existing Big Data analytics processing methods and models. Hadoop® 2 Quick-Start Guide is the first easy, accessible guide to Apache Hadoop 2.x, YARN, and the modern Hadoop ecosystem. Building on his unsurpassed experience teaching Hadoop and Big Data, author Douglas Eadline covers all the basics you need to know to install and use Hadoop 2 on personal computers or servers, and to navigate the powerful technologies that complement it. Eadline concisely introduces and explains every key Hadoop 2 concept, tool, and service, illustrating each with a simple "beginning-to-end" example and identifying trustworthy, up-to-date resources for learning more. This guide is ideal if you want to learn about Hadoop 2 without getting mired in technical details. Douglas Eadline will bring you up to speed quickly, whether you're a user, admin, devops specialist, programmer, architect, analyst, or data scientist. Coverage Includes Understanding what Hadoop 2 and YARN do, and how they improve on Hadoop 1 with MapReduce Understanding Hadoop-based Data Lakes versus RDBMS Data Warehouses Installing Hadoop 2 and core services on Linux machines, virtualized sandboxes, or clusters Exploring the Hadoop Distributed File System (HDFS) Understanding the essentials of MapReduce and YARN application programming Simplifying programming and data movement with Apache Pig, Hive, Sqoop, Flume, Oozie, and HBase Observing application progress, controlling jobs, and managing workflows Managing Hadoop efficiently with Apache Ambari—including recipes for HDFS to NFSv3 gateway, HDFS snapshots, and YARN configuration Learning basic Hadoop 2 troubleshooting, and installing Apache Hue and Apache Spark

Big Data in Practice Bernard Marr 2016-03-22 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

Programming Hive Edward Capriolo 2012-09-26 Describes the features and functions of Apache Hive, the data infrastructure for Hadoop.

Disruptive Possibilities: How Big Data Changes Everything Jeffrey Needham 2013-05-06 Big data has more disruptive potential than any information technology developed in the past 40 years. As author Jeffrey Needham points out in this revealing book, big data can provide unprecedented visibility into the operational efficiency of enterprises and agencies. Disruptive Possibilities provides an historically-informed overview through a wide range of topics, from the evolution of commodity supercomputing and the simplicity of big data technology, to the ways

conventional clouds differ from Hadoop analytics clouds. This relentlessly innovative form of computing will soon become standard practice for organizations of any size attempting to derive insight from the tsunami of data engulfing them. Replacing legacy silos—whether they're infrastructure, organizational, or vendor silos—with a platform-centric perspective is just one of the big stories of big data. To reap maximum value from the myriad forms of data, organizations and vendors will have to adopt highly collaborative habits and methodologies.

Big Data Analytics: Systems, Algorithms, Applications C.S.R. Prabhu 2019-10-14 This book provides a comprehensive survey of techniques, technologies and applications of Big Data and its analysis. The Big Data phenomenon is increasingly impacting all sectors of business and industry, producing an emerging new information ecosystem. On the applications front, the book offers detailed descriptions of various application areas for Big Data Analytics in the important domains of Social Semantic Web Mining, Banking and Financial Services, Capital Markets, Insurance, Advertisement, Recommendation Systems, Bio-Informatics, the IoT and Fog Computing, before delving into issues of security and privacy. With regard to machine learning techniques, the book presents all the standard algorithms for learning — including supervised, semi-supervised and unsupervised techniques such as clustering and reinforcement learning techniques to perform collective Deep Learning. Multi-layered and nonlinear learning for Big Data are also covered. In turn, the book highlights real-life case studies on successful implementations of Big Data Analytics at large IT companies such as Google, Facebook, LinkedIn and Microsoft. Multi-sectorial case studies on domain-based companies such as Deutsche Bank, the power provider Opower, Delta Airlines and a Chinese City Transportation application represent a valuable addition. Given its comprehensive coverage of Big Data Analytics, the book offers a unique resource for undergraduate and graduate students, researchers, educators and IT professionals alike.

Big Data Analytics with Spark Mohammed Guller 2015-12-29 Big Data Analytics with Spark is a step-by-step guide for learning Spark, which is an open-source fast and general-purpose cluster computing framework for large-scale data analysis. You will learn how to use Spark for different types of big data analytics projects, including batch, interactive, graph, and stream data analysis as well as machine learning. In addition, this book will help you become a much sought-after Spark expert. Spark is one of the hottest Big Data technologies. The amount of data generated today by devices, applications and users is exploding. Therefore, there is a critical need for tools that can analyze large-scale data and unlock value from it. Spark is a powerful technology that meets that need. You can, for example, use Spark to perform low latency computations through the use of efficient caching and iterative algorithms; leverage the features of its shell for easy and interactive Data analysis; employ its fast batch processing and low latency features to process your real time data streams and so on. As a result, adoption of Spark is rapidly growing and is replacing Hadoop MapReduce as the technology of choice for big data analytics. This book provides an introduction to Spark and related big-data technologies. It covers Spark core and its add-on libraries, including Spark SQL, Spark Streaming, GraphX, and MLlib. Big Data Analytics with Spark is therefore written for busy professionals who prefer learning a new technology from a consolidated source instead of spending countless hours on the Internet trying to pick bits and pieces from different sources. The book also provides a chapter on Scala, the hottest functional programming language, and the program that underlies Spark. You'll learn the basics of

functional programming in Scala, so that you can write Spark applications in it. What's more, Big Data Analytics with Spark provides an introduction to other big data technologies that are commonly used along with Spark, like Hive, Avro, Kafka and so on. So the book is self-sufficient; all the technologies that you need to know to use Spark are covered. The only thing that you are expected to know is programming in any language. There is a critical shortage of people with big data expertise, so companies are willing to pay top dollar for people with skills in areas like Spark and Scala. So reading this book and absorbing its principles will provide a boost—possibly a big boost—to your career.

Big Data, Analytics, and the Future of Marketing & Sales McKinsey Chief McKinsey Chief Marketing & Sales Officer Forum 2014-08-16 Big Data is the biggest game-changing opportunity for marketing and sales since the Internet went mainstream almost 20 years ago. The data big bang has unleashed torrents of terabytes about everything from customer behaviors to weather patterns to demographic consumer shifts in emerging markets. This collection of articles, videos, interviews, and slideshares highlights the most important lessons for companies looking to turn data into above-market growth: Using analytics to identify valuable business opportunities from the data to drive decisions and improve marketing return on investment (MROI) Turning those insights into well-designed products and offers that delight customers Delivering those products and offers effectively to the marketplace.The goldmine of data represents a pivot-point moment for marketing and sales leaders. Companies that inject big data and analytics into their operations show productivity rates and profitability that are 5 percent to 6 percent higher than those of their peers. That's an advantage no company can afford to ignore.