Apache Mahout Beyond Mapreduce

This book constitutes the workshop proceedings of the 18th International Conference on Database Systems for Advanced Applications, DASFAA 2013, held in Wuhan, China, in April 2013. The volume contains three workshops, each focusing on specific area that contributes to the main themes of the DASFAA conference: The First International Workshop on Big Data Management and Analytics (BDMA 2013), the Third International Workshop on Social Networks and Social Web (SNSM 2013) and the International Workshop on Semantic Computing and Personalization (SeCoP 2013).

There is an easier way to build Hadoop applications. With this hands-on book, you'll learn how to use Cascading, the open source abstraction framework for Hadoop that lets you easily create and manage powerful enterprise-grade data processing applications—without having to learn the intricacies of MapReduce. Working with sample apps based on Java and other JVM languages, you'll quickly learn Cascading's streamlined approach to data processing, data filtering, and workflow optimization. This book demonstrates how this framework can help your business extract meaningful information from large amounts of distributed data. Start working on Cascading example projects right away Model and analyze unstructured data in any format, from any source Build and test applications with familiar constructs and reusable components Work with the Scalding and Cascalog Domain-Specific Languages Easily deploy applications to Hadoop, regardless of cluster location or data size Build workflows that integrate several big data frameworks and processes Explore common use cases for Cascading, including features and tools that support them Examine a case study that uses a dataset from the Open Data Initiative

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.

If your team is investigating ways to design applications for the cloud, this concise book introduces 11 architecture patterns that can help you take advantage of cloud-platform services. You'll learn how each of these platform-agnostic patterns work, when they might be useful in the cloud, and what impact they'll have on your application architecture. You'll also see an example of each pattern applied to an application built with Windows Azure. The patterns are organized into four major topics, such as scalability and handling failure, and primer chapters provide background on each topic. With the information in this book, you'll be able to make informed decisions for designing effective cloud-native applications that maximize the value of cloud services, while also paying attention to user experience and operational efficiency. Learn about architectural patterns for: Scalability. Discover the advantages of horizontal scaling. Patterns covered include Horizontally Scaling Compute, Queue-Centric Workflow, and Auto-Scaling. Big data. Learn how to handle large amounts of data across a distributed system. Eventual consistency is explained, along with the MapReduce and Database Sharding patterns. Handling failure. Understand how multitenant cloud services and commodity hardware influence your applications. Patterns covered include Busy Signal and Node Failure. Distributed users. Learn how to overcome delays due to network latency when building applications for a geographically distributed user base. Patterns covered include Colocation, Valet Key, CDN, and Multi-Site Deployment.

This book gathers selected high-quality papers presented at the International Conference on Computing, Power and Communication Technologies 2019 (GUCON 2019), organized by Galgotias University, India, in September 2019. The content is divided into three sections – data mining and big data analysis, communication technologies, and cloud computing and computer networks. In-depth discussions of various issues within these broad areas provide an intriguing and insightful reference guide for researchers, engineers and students alike.

Apache MahoutBeyond Mapreduce

Ongoing advancements in modern technology have led to significant developments in artificial intelligence. With the numerous applications available, it becomes imperative to conduct research and make further progress in this field. Artificial Intelligence: Concepts, Methodologies, Tools, and Applications provides a comprehensive overview of the latest breakthroughs and recent progress in artificial intelligence. Highlighting relevant technologies, uses, and techniques across various industries and settings, this publication is a pivotal reference source for researchers, professionals, academics, upper-level students, and practitioners interested in emerging perspectives in the field of artificial intelligence.

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

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

If you're training a machine learning model but aren't sure how to put it into production, this book will get you there. Kubeflow provides a collection of cloud native tools for different stages of a model's lifecycle, from data exploration, feature preparation, and model training to model serving. This guide helps data scientists build production-grade machine learning implementations with Kubeflow and shows data engineers how to make models scalable and reliable. Using examples throughout the book, authors Holden Karau, Trevor Grant, Ilan Filonenko, Richard Liu, and Boris Lublinsky explain how to use Kubeflow to train and serve your machine learning models on top of Kubernetes in the cloud or in a development environment on-premises. Understand Kubeflow's design, core components, and the problems it solves Understand the differences between Kubeflow on different cluster types Train models using Kubeflow with popular tools including Scikit-learn, TensorFlow, and Apache Spark Keep your model up to date with Kubeflow Pipelines Understand how to capture model training metadata Explore how to extend Kubeflow with additional open source tools Use hyperparameter tuning for training Learn how to serve your model in production

There’s growing interest in learning how to analyze streaming data in large-scale systems such as web traffic, financial transactions, machine logs, industrial sensors, and many others. But analyzing data streams at scale has been difficult to do well—until now. This practical book delivers a deep introduction to Apache Flink, a highly innovative open source stream processor with a surprising range of capabilities. Authors Ellen Friedman and Kostas Tzoumas show technical and nontechnical readers alike how Flink is engineered to overcome significant tradeoffs that have limited the effectiveness of other approaches to stream processing. You’ll also learn how Flink has the ability to handle both stream and batch data processing with one technology. Learn the consequences of not doing streaming well—in retail and marketing, IoT, telecom, and banking and finance Explore how to design data architecture to gain the best advantage from stream processing Get an overview of Flink’s capabilities and features, along with examples of how companies use Flink, including in production Take a technical dive into Flink, and learn how it handles time and stateful computation Examine how Flink processes both streaming (unbounded) and batch (bounded) data without sacrificing performance

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.

Apache Mahout is a scalable machine learning library with algorithms for clustering, classification, and recommendations. It empowers users to analyze patterns in large, diverse, and complex datasets faster and more scalably. This book is an all-inclusive guide to analyzing large and complex datasets using Apache Mahout. It explains complicated but very effective machine learning algorithms simply, in relation to real-world practical examples. Starting from the fundamental concepts of machine learning and Apache Mahout, this book guides you through Apache Mahout's implementations of machine learning techniques including classification, clustering, and recommendations. During this exciting walkthrough, real-world applications, a diverse range of popular algorithms and their implementations, code examples, evaluation strategies, and best practices are given for each technique. Finally, you will learn vdata visualization techniques for Apache Mahout to bring your data to life.

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

Engineering education methods and standards are important features of engineering programs that should be carefully designed both to provide students and stakeholders with valuable, active, integrated learning experiences, and to provide a vehicle for assessing program outcomes. With the driving force of the globalization of the engineering profession, standards should be developed for mutual recognition of engineering education across the world, but it is proving difficult to achieve. The Handbook of Research on Engineering Education in a Global Context provides innovative insights into the importance of quality training and preparation for engineering students. It explores the common and current problems encountered in areas such as quality and standards, management information systems, innovation and enhanced learning technologies in education, as well as the challenges of employability, entrepreneurship, and diversity. This publication is vital reference source for science and engineering educators, engineering professionals, and educational administrators interested in topics centered on the education of students in the field of engineering.

This handbook offers comprehensive coverage of recent advancements in Big Data technologies and related paradigms. Chapters are authored by international leading experts in the field, and have been reviewed and revised for maximum reader value. The volume consists of twenty-five chapters organized into four main parts. Part one covers the fundamental concepts of Big Data technologies including data curation mechanisms, data models, storage models, programming models and programming platforms. It also dives into the details of implementing Big SQL query engines and big stream processing systems. Part Two focuses on the semantic aspects of Big Data management including data integration and exploratory ad hoc analysis in addition to structured querying and pattern matching techniques. Part Three presents a comprehensive overview of large scale graph processing. It covers the most recent research in large scale graph processing platforms, introducing several scalable graph querying and mining mechanisms in domains such as social networks. Part Four details novel applications that have been made possible by the rapid emergence of Big Data technologies such as Internet-of-Things (IOT), Cognitive Computing and SCADA Systems. All parts of the book discuss open research problems, including potential opportunities, that have arisen from the rapid progress of Big Data technologies and the associated increasing requirements of application domains. Designed for researchers, IT professionals and graduate students, this book is a timely contribution to the growing Big Data field. Big Data has been recognized as one of leading emerging technologies that will have a major contribution and impact on the various fields of science and varies aspect of the human society over the coming decades. Therefore, the content in this book will be an essential tool to help readers understand the development and future of the field.

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

"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

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

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

If you are a system or application developer interested in learning how to solve practical problems using the Hadoop framework, then this book is ideal for you. You are expected to be familiar with the Unix/Linux command-line interface and have some experience with the Java programming language. Familiarity with Hadoop would be a plus.

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

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

This is the first book to offer a comprehensive yet concise overview of the challenges and opportunities presented by the use of big data in healthcare. The respective chapters address a range of aspects: from health management to patient safety; from the human factor perspective to ethical and economic considerations, and many more. By providing a historical background on the use of big data, and critically analyzing current approaches together with issues and challenges related to their applications, the book not only sheds light on the problems entailed by big data, but also paves the way for possible solutions and future research directions. Accordingly, it offers an insightful reference guide for health information technology professionals, healthcare managers, healthcare practitioners, and patients alike, aiding them in their decision-making processes; and for students and researchers whose work involves data science-related research issues in healthcare.

This book provides a general and comprehensible overview of imbalanced learning. It contains a formal description of a problem, and focuses on its main features, and the most relevant proposed solutions. Additionally, it considers the different scenarios in Data Science for which the imbalanced classification can create a real challenge. This book stresses the gap with standard classification tasks by reviewing the case studies and ad-hoc performance metrics that are applied in this area. It also covers the different approaches that have been traditionally applied to address the binary skewed class distribution. Specifically, it reviews cost-sensitive learning, data-level preprocessing methods and algorithm-level solutions, taking also into account those ensemble-learning solutions that embed any of the former alternatives. Furthermore, it focuses on the extension of the problem for multi-class problems, where the former classical methods are no longer to be applied in a straightforward way. This book also focuses on the data intrinsic characteristics that are the main causes which, added to the uneven class distribution, truly hinders the performance of classification algorithms in this scenario. Then, some notes on data reduction are provided in order to understand the advantages related to the use of this type of approaches. Finally this book introduces some novel areas of study that are gathering a deeper attention on the imbalanced data issue. Specifically, it considers the classification of data streams, non-classical classification problems, and the scalability related to Big Data. Examples of software libraries and modules to address imbalanced classification are provided. This book is highly suitable for technical professionals, senior undergraduate and graduate students in the areas of data science, computer science and engineering. It will also be useful for scientists and researchers to gain insight on the current developments in this area of study, as well as future research directions.

Machine Learning and the Internet of Medical Things in Healthcare discusses the applications and challenges of machine learning for healthcare applications. The book provides a platform for presenting machine learning-enabled healthcare techniques and offers a mathematical and conceptual background of the latest technology. It describes machine learning techniques along with the emerging platform of the Internet of Medical Things used by practitioners and researchers worldwide. The book includes deep feed forward networks, regularization, optimization algorithms, convolutional networks, sequence modeling, and practical methodology. It also presents the concepts of the Internet of Things, the set of technologies that develops traditional devices into smart devices. Finally, the book offers research perspectives, covering the convergence of machine learning and IoT. It also presents the application of these technologies in the development of healthcare frameworks. Provides an introduction to the Internet of Medical Things through the principles and applications of machine learning Explains the functions and applications of machine learning in various applications such as ultrasound imaging, biomedical signal processing, robotics, and biomechatronics Includes coverage of the evolution of healthcare applications with machine learning, including Clinical Decision Support Systems, artificial intelligence in biomedical engineering, and AI-enabled connected health informatics, supported by real-world case studies

This book constitutes the proceedings of the International Conference on E-business and Strategy, iCETS 2012, held in Tianjin, China, in August 2012. The 65 revised full papers presented were carefully reviewed and selected from 231 submissions. The papers feature contemporary research on developments in the fields of e-business technology, information management systems, and business strategy. Topics addressed are latest development on e-business technology, computer science and software engineering for e-business, e-business and e-commerce applications, social networking and social engineering for e-business, e-business strategic management and economics development, e-business education, entrepreneurship and e-learning, digital economy strategy, as well as internet and e-commerce policy.

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

edge model-driven analytics. Big Data Analytics Beyond Hadoop is an indispensable resource for everyone who wants to reach the cutting edge of Big Data analytics, and stay there: practitioners, architects, programmers, data scientists, researchers, startup entrepreneurs, and advanced students.

Production-targeted Spark guidance with real-world use cases Spark: Big Data Cluster Computing in Production goes beyond general Spark overviews to provide targeted guidance toward using lightning-fast big-data clustering in production. Written by an expert team well-known in the big data community, this book walks you through the challenges in moving from proof-of-concept or demo Spark applications to live Spark in production. Real use cases provide deep insight into common problems, limitations, challenges, and opportunities, while expert tips and tricks help you get the most out of Spark performance. Coverage includes Spark SQL, Tachyon, Kerberos, ML Lib, YARN, and Mesos, with clear, actionable guidance on resource scheduling, db connectors, streaming, security, and much more. Spark has become the tool of choice for many Big Data problems, with more active contributors than any other Apache Software project. General introductory books abound, but this book is the first to provide deep insight and real-world advice on using Spark in production. Specific guidance, expert tips, and invaluable foresight make this guide an incredibly useful resource for real production settings. Review Spark hardware requirements and estimate cluster size Gain insight from real-world production use cases Tighten security, schedule resources, and fine-tune performance Overcome common problems encountered using Spark in production Spark works with other big data tools including MapReduce and Hadoop, and uses languages you already know like Java, Scala, Python, and R. Lightning speed makes Spark too good to pass up, but understanding limitations and challenges in advance goes a long way toward easing actual production implementation. Spark: Big Data Cluster Computing in Production tells you everything you need to know, with real-world production insight and expert guidance, tips, and tricks.

The volume of data that is generated, stored, and communicated across different industrial sections, business units, and scientific research communities has been rapidly expanding. The recent developments in cellular telecommunications and distributed/parallel computation technology have enabled real-time collection and processing of the generated data across different sections. On the one hand, the internet of things (IoT) enabled by cellular telecommunication industry connects various types of sensors that can collect heterogeneous data. On the other hand, the recent advances in computational capabilities such as parallel processing in graphical processing units (GPUs) and distributed processing over cloud computing clusters enabled the processing of a vast amount of data. There has been a vital need to discover important patterns and infer trends from a large volume of data (so-called Big Data) to empower data-driven decision-making processes. Tools and techniques have been developed in machine learning to draw insightful conclusions from available data in a structured and automated fashion. Machine learning algorithms are based on concepts and tools developed in several fields including statistics, artificial intelligence, information theory, cognitive science, and control theory. The recent advances in machine learning have had a broad range of applications in different scientific disciplines. This book covers recent advances of machine learning techniques in a broad range of applications in smart cities, automated industry, and emerging businesses.

Social networking has increased drastically in recent years, resulting in an increased amount of data being created daily. Furthermore, diversity of issues and complexity of the social networks pose a challenge in social network mining. Traditional algorithm software cannot deal with such complex and vast amounts of data, necessitating the development of novel analytic approaches and tools. This reference work deals with social network aspects of big data analytics. It covers theory, practices and challenges in social networking. The book spans numerous disciplines like neural networking, deep learning, artificial intelligence, visualization, e-learning in higher education, e-healthcare, security and intrusion detection.

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.

Learn advanced analytical techniques and leverage existing tool kits to make your analytic applications more powerful, precise, and efficient. This book provides the right combination of architecture, design, and implementation information to create analytical systems that go beyond the basics of classification, clustering, and recommendation. Pro Hadoop Data Analytics emphasizes best practices to ensure coherent, efficient development. A complete example system will be developed using standard third-party components that consist of the tool kits, libraries, visualization and reporting code, as well as support glue to provide a working and extensible end-to-end system. The book also highlights the importance of end-to-end, flexible, configurable, high-performance data pipeline systems with analytical components as well as appropriate visualization results. You'll discover the importance of mix-and-match or hybrid systems, using different analytical components in one application. This hybrid approach will be prominent in the examples. What You'll Learn Build big data analytic systems with the Hadoop ecosystem Use libraries, tool kits, and algorithms to make development easier and more effective Apply metrics to measure performance and efficiency of components and systems Connect to standard relational databases, noSQL data sources, and more Follow case studies with example components to create your own systems Who This Book Is For Software engineers, architects, and data scientists with an interest in the design and implementation of big data analytical systems using Hadoop, the Hadoop ecosystem, and other associated technologies.

This book is the basic guide for developers, architects, engineers, and anyone who wants to start leveraging the open-source software Hadoop and Hive to build distributed, scalable concurrent big data applications. Hive will be used for reading, writing, and managing the large, data set files. The book is a concise guide on getting started with an overall understanding on Apache Hadoop and Hive and how they work together to speed up development with minimal effort. It will refer to simple concepts and examples, as they are likely to be the best teaching aids. It will explain the logic, code, and configurations needed to build a successful, distributed, concurrent application, as well as the

reason behind those decisions. FEATURES: Shows how to leverage the open-source software Hadoop and Hive to build distributed, scalable, concurrent big data applications Includes material on Hive architecture with various storage types and the Hive query language Features a chapter on big data and how Hadoop can be used to solve the changes around it Explains the basic Hadoop setup, configuration, and optimization

This volume helps to fill the gap between data analytics, image processing, and soft computing practices. Soft computing methods are used to focus on data analytics and image processing to develop good intelligent systems. To this end, readers of this volume will find quality research that presents the current trends, advanced methods, and hybridized techniques relating to data analytics and intelligent systems. The book also features case studies related to medical diagnosis with the use of image processing and soft computing algorithms in particular models. Providing extensive coverage of biometric systems, soft computing, image processing, artificial intelligence, and data analytics, the chapter authors discuss the latest research issues, present solutions to research problems, and look at comparative analysis with earlier results. Topics include some of the most important challenges and discoveries in intelligent systems today, such as computer vision concepts and image identification, data analysis and computational paradigms, deep learning techniques, face and speaker recognition systems, and more.

The growth of a global digital economy has enabled rapid communication, instantaneous movement of funds, and availability of vast amounts of information. With this come challenges such as the vulnerability of digitalized sociotechnological systems (STSs) to destructive events (earthquakes, disease events, terrorist attacks). Similar issues arise for disruptions to complex linked natural and social systems (from changing climates, evolving urban environments, etc.). This book explores new approaches to the resilience of sociotechnological and natural-social systems in a digital world of big data, extraordinary computing capacity, and rapidly developing methods of Artificial Intelligence. Most of the book’s papers were presented at the Workshop on Big Data and Systems Analysis held at the International Institute for Applied Systems Analysis in Laxenburg, Austria in February, 2020. Their authors are associated with the Task Group “Advanced mathematical tools for data-driven applied systems analysis” created and sponsored by CODATA in November, 2018. The world-wide COVID-19 pandemic illustrates the vulnerability of our healthcare systems, supply chains, and social infrastructure, and confronts our notions of what makes a system resilient. We have found that use of AI tools can lead to problems when unexpected events occur. On the other hand, the vast amounts of data available from sensors, satellite images, social media, etc. can also be used to make modern systems more resilient. Papers in the book explore disruptions of complex networks and algorithms that minimize departure from a previous state after a disruption; introduce a multigrammatical framework for the technological and resource bases of today’s large-scale industrial systems and the transformations resulting from disruptive events; and explain how robotics can enhance pre-emptive measures or post-disaster responses to increase resiliency. Other papers explore current directions in data processing and handling and principles of FAIRness in data; how the availability of large amounts of data can aid in the development of resilient STSs and challenges to overcome in doing so. The book also addresses interactions between humans and built environments, focusing on how AI can inform today’s smart and connected buildings and make them resilient, and how AI tools can increase resilience to misinformation and its dissemination.

The volume LNCS 12393 constitutes the papers of the 22nd International Conference Big Data Analytics and Knowledge Discovery which will be held online in September 2020. The 15 full papers presented together with 14 short papers plus 1 position paper in this volume were carefully reviewed and selected from a total of 77 submissions. This volume offers a wide range to following subjects on theoretical and practical aspects of big data analytics and knowledge discovery as a new generation of big data repository, data pre-processing, data mining, text mining, sequences, graph mining, and parallel processing.

Summary Mahout in Action is a hands-on introduction to machine learning with Apache Mahout. Following real-world examples, the book presents practical use cases and then illustrates how Mahout can be applied to solve them. Includes a free audio- and video-enhanced ebook. About the Technology A computer system that learns and adapts as it collects data can be really powerful. Mahout, Apache's open source machine learning project, captures the core algorithms of recommendation systems, classification, and clustering in ready-to-use, scalable libraries. With Mahout, you can immediately apply to your own projects the machine learning techniques that drive Amazon, Netflix, and others. About this Book This book covers machine learning using Apache Mahout. Based on experience with real-world applications, it introduces practical use cases and illustrates how Mahout can be applied to solve them. It places particular focus on issues of scalability and how to apply these techniques against large data sets using the Apache Hadoop framework. This book is written for developers familiar with Java -- no prior experience with Mahout is assumed. Owners of a Manning pBook purchased anywhere in the world can download a free eBook from manning.com at any time. They can do so multiple times and in any or all formats available (PDF, ePub or Kindle). To do so, customers must register their printed copy on Manning's site by creating a user account and then following instructions printed on the pBook registration insert at the front of the book. What's Inside Use group data to make individual recommendations Find logical clusters within your data Filter and refine with on-the-fly classification Free audio and video extras Table of Contents Meet Apache Mahout PART 1 RECOMMENDATIONS Introducing recommenders Representing recommender data Making recommendations Taking recommenders to production Distributing recommendation computations PART 2 CLUSTERING Introduction to clustering Representing data Clustering algorithms in Mahout Evaluating and improving clustering quality Taking clustering to production Real-world applications of clustering PART 3 CLASSIFICATION Introduction to classification Training a classifier Evaluating and tuning a classifier Deploying a classifier Case study: Shop It To Me

A hands-on guide to leveraging NoSQL databases NoSQL databases are an efficient and powerful tool for storing and manipulating vast quantities of data. Most NoSQL

databases scale well as data grows. In addition, they are often malleable and flexible enough to accommodate semi-structured and sparse data sets. This comprehensive hands-on guide presents fundamental concepts and practical solutions for getting you ready to use NoSQL databases. Expert author Shashank Tiwari begins with a helpful introduction on the subject of NoSQL, explains its characteristics and typical uses, and looks at where it fits in the application stack. Unique insights help you choose which NoSQL solutions are best for solving your specific data storage needs. Professional NoSQL: Demystifies the concepts that relate to NoSQL databases, including column-family oriented stores, key/value databases, and document databases. Delves into installing and configuring a number of NoSQL products and the Hadoop family of products. Explains ways of storing, accessing, and querying data in NoSQL databases through examples that use MongoDB, HBase, Cassandra, Redis, CouchDB, Google App Engine Datastore and more. Looks at architecture and internals. Provides guidelines for optimal usage, performance tuning, and scalable configurations. Presents a number of tools and utilities relating to NoSQL, distributed platforms, and scalable processing, including Hive, Pig, RRDtool, Nagios, and more.

If your organization is about to enter the world of big data, you not only need to decide whether Apache Hadoop is the right platform to use, but also which of its many components are best suited to your task. This field guide makes the exercise manageable by breaking down the Hadoop ecosystem into short, digestible sections. You'll quickly understand how Hadoop's projects, subprojects, and related technologies work together. Each chapter introduces a different topic—such as core technologies or data transfer—and explains why certain components may or may not be useful for particular needs. When it comes to data, Hadoop is a whole new ballgame, but with this handy reference, you'll have a good grasp of the playing field. Topics include: Core technologies—Hadoop Distributed File System (HDFS), MapReduce, YARN, and Spark Database and data management—Cassandra, HBase, MongoDB, and Hive Serialization—Avro, JSON, and Parquet Management and monitoring—Puppet, Chef, Zookeeper, and Oozie Analytic helpers—Pig, Mahout, and MLLib Data transfer—Scoop, Flume, distcp, and Storm Security, access control, auditing—Sentry, Kerberos, and Knox Cloud computing and virtualization—Serengeti, Docker, and Whirr