Apache Hadoop 2 Yarn Best Practices In The Apache Hadoop Ecosystem

Learn to build powerful machine learning models quickly and deploy large-scale predictive applications About This Book Design, engineer and deploy scalable machine learning solutions with the power of Python Take command of Hadoop and Spark with Python for effective machine learning on a map reduce framework Build state-of-the-art models and develop personalized recommendations to perform machine learning at scale Who This Book Is For This book is for anyone who intends to work with large and complex data sets. Familiarity with basic Python and machine learning concepts is recommended. Working knowledge in statistics and computational mathematics would also be helpful. What You Will Learn Apply the most scalable machine learning algorithms Work with modern state-of-the-art large-scale machine learning techniques Increase predictive accuracy with deep learning and scalable data-handling techniques Improve your work by combining the MapReduce framework with Spark Build powerful ensembles at scale Use data streams to train linear and non-linear predictive models from extremely large datasets using a single machine In Detail Large Python machine learning projects involve new problems associated with specialized machine learning architectures and designs that many data scientists have yet to tackle. But finding algorithms and designing and building platforms that deal with large sets of data is a growing need. Data scientists have to manage and maintain increasingly complex data projects, and with the rise of big data comes an increasing demand for computational and algorithmic efficiency. Large Scale Machine Learning with Python uncovers a new wave of machine learning algorithms that meet scalability demands together with a high predictive accuracy. Dive into scalable machine learning and the three forms of scalability. Speed up algorithms that can be used on a desktop computer with tips on parallelization and memory allocation. Get to grips with new algorithms that are specifically designed for large projects and can handle bigger files, and learn about machine learning in big data environments. We will also cover the most effective machine learning techniques on a map reduce framework in Hadoop and Spark in Python. Style and Approach This efficient and practical title is stuffed full of the techniques, tips and tools you need to ensure your large scale Python machine learning runs swiftly and seamlessly. Large-scale machine learning tackles a different issue to what is currently on the market. Those working with Hadoop clusters and in data intensive environments can now learn effective ways of building powerful machine learning models from prototype to production. This book is written in a style that programmers from other languages (R, Julia, Java, Matlab) can follow.

This book constitutes the refereed proceedings of the 16th International Conference on Algorithms and Architectures for Parallel Processing, ICA3PP 2016, held in Granada, Spain, in December 2016. The 30 full papers and 22 short papers presented were carefully reviewed and selected from 117 submissions. They cover many dimensions of parallel algorithms and architectures, encompassing fundamental theoretical approaches, practical experimental projects, and commercial components and systems trying to push beyond the limits of existing technologies, including experimental efforts, innovative systems, and investigations that identify weaknesses in existing parallel processing technology.

Practical Graph Analytics with Apache Giraph helps you build data mining and machine learning applications using the Apache Foundation’s Giraph framework for graph processing. This is the same framework as used by Facebook, Google, and other social media analytics operations to derive business value from vast amounts of interconnected data points. Graphs arise in a wealth of data scenarios and describe the connections that are naturally formed in both digital and real worlds. Examples of such connections abound in online social networks such as Facebook and Twitter, among users who rate movies from services like Netflix and Amazon Prime, and are useful even in the context of biological networks for scientific research. Whether in the context of business or science, viewing data as connected adds value by increasing the amount of information available to be drawn from that data and put to use in generating new revenue or scientific opportunities. Apache Giraph offers a simple yet flexible programming model targeted to graph algorithms and designed to scale easily to accommodate massive amounts of data. Originally developed at Yahoo!, Giraph is now a top top-level project at the Apache Foundation, and it enlists contributors from companies such as Facebook, LinkedIn, and Twitter. Practical Graph Analytics with Apache Giraph brings the power of Apache Giraph to you, showing how to harness the power of graph processing for your own data by building sophisticated graph analytics applications using the very same framework that is relied upon by some of the largest players in the industry today.

This book presents a compilation of current trends, technologies, and challenges in connection with Big Data. Many fields of science and engineering are data-driven, or generate huge amounts of data that are ripe for the picking. There are now more sources of data than ever before, and more means of capturing data. At the same time, the sheer volume and complexity of the data have sparked new developments, where many Big Data problems require new solutions. Given its scope, the book offers a valuable reference guide for all graduate students, researchers, and scientists interested in exploring the potential of Big Data applications.

Job Scheduling Strategies for Parallel Processing

Handbook of IoT and Big Data

Hadoop Real-World Solutions Cookbook

Learn the Essentials of Big Data Computing in the Apache Hadoop 2 Ecosystem

Big Data with Hadoop MapReduce

Data, Engineering and Applications

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

Pro Apache Hadoop, Second Edition brings you up to speed on Hadoop – the framework of big data. Revised to cover Hadoop 2.0, the book covers the very latest developments such as YARN (aka MapReduce 2.0), new HDFS high-availability features, and increased scalability in the form of HDFS Federations. All the old content has been revised too, giving the latest on the ins and outs of MapReduce, cluster design, the Hadoop Distributed File System, and more. This book covers everything you need to build your first Hadoop cluster and begin analyzing and deriving value from your business and scientific data. Learn to solve big-data problems the MapReduce way, by breaking a big problem into chunks and creating small-scale solutions that can be flung across thousands upon thousands of nodes to analyze large data volumes in a short amount of wall-clock time. Learn how to let Hadoop take care of distributing and parallelizing your software—you just focus on the code; Hadoop takes care of the rest. Covers all that is new in Hadoop 2.0 Written by a professional involved in Hadoop since day one Takes you quickly to the seasoned pro level on the hottest cloud-computing framework

Over 90 hands-on recipes to help you learn and master the intricacies of Apache Hadoop 2.X, YARN, Hive, Pig, Oozie, Flume, Sqoop, Apache Spark, and Mahout About This Book Implement outstanding Machine Learning use cases on your own analytics models and processes. Solutions to common problems when working with the Hadoop ecosystem. Step-by-step implementation of end-to-end big data use cases. Who This Book Is For Readers who have a basic knowledge of big data systems and want to advance their knowledge with hands-on recipes. What You Will Learn Installing and maintaining Hadoop 2.X cluster and its ecosystem. Write advanced Map Reduce programs and understand design patterns. Advanced Data Analysis using the Hive, Pig, and Map Reduce programs. Import and export data from various sources using Sqoop and Flume. Data storage in various file formats such as Text, Sequential, Parquet, ORC, and RC Files. Machine learning principles with libraries such as Mahout Batch and Stream data processing using Apache Spark In Detail Big data is the current requirement. Most organizations produce huge amount of data every day. With the arrival of Hadoop-like tools, it has become easier for everyone to solve big data problems with great efficiency and at minimal cost. Grasping Machine Learning techniques will help you greatly in building predictive models and using this data to make the right decisions for your organization. Hadoop Real World Solutions Cookbook gives readers insights into learning and mastering big data via recipes. The book not only clarifies most big data tools in the market but also provides best practices for using them. The book provides recipes that are based on the latest versions of Apache Hadoop 2.X, YARN, Hive, Pig, Sqoop, Flume, Apache Spark, Mahout and many more such ecosystem tools. This real-world-solution cookbook is packed with handy recipes you can apply to your own everyday issues. Each chapter provides in-depth recipes that can be referenced easily. This book provides detailed practices on the latest technologies such as YARN and Apache Spark. Readers will be able to consider themselves as big data experts on completion of this book. This guide is an invaluable tutorial if you are planning to implement a big data warehouse for your business. Style and approach An easy-to-follow guide that walks you through world of big data. Each tool in the Hadoop ecosystem is explained in detail and the recipes are placed in such a manner that readers can implement them sequentially. Plenty of reference links are provided for advanced reading.

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 First Step towards Hadoop Administration and Management

Pro Apache Hadoop

Challenges, Solutions and Perspectives

Research Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing

16th International Conference, ICA3PP 2016, Granada, Spain, December 14-16, 2016, Proceedings

Apache Spark 2.x Cookbook

If you have a working knowledge of Hadoop 1.x but want to start afresh with YARN, this book is ideal for you. You will be able to install and administer a YARN cluster and also discover the configuration settings to fine-tune your cluster both in terms of performance and scalability. This book will help you develop, deploy, and run multiple applications/frameworks on the same shared YARN cluster.

Construct a robust end-to-end solution for analyzing and visualizing streaming data Real-time analytics is the hottest topic in data analytics today. In Real-Time Analytics: Techniques to Analyze

and Visualize Streaming Data, expert Byron Ellis teaches data analysts technologies to build an effective real-time analytics platform. This platform can then be used to make sense of the constantly changing data that is beginning to outpace traditional batch-based analysis platforms. The author is among a very few leading experts in the field. He has a prestigious background in research, development, analytics, real-time visualization, and Big Data streaming and is uniquely qualified to help you explore this revolutionary field. Moving from a description of the overall analytic architecture of real-time analytics to using specific tools to obtain targeted results, Real-Time Analytics leverages open source and modern commercial tools to construct robust, efficient systems that can provide real-time analysis in a cost-effective manner. The book includes: A deep discussion of streaming data systems and architectures Instructions for analyzing, storing, and delivering streaming data Tips on aggregating data and working with sets Information on data warehousing options and techniques Real-Time Analytics includes in-depth case studies for website analytics, Big Data, visualizing streaming and mobile data, and mining and visualizing operational data flows. The book's "recipe" layout lets readers quickly learn and implement different techniques. All of the code examples presented in the book, along with their related data sets, are available on the companion website.

As more corporations turn to Hadoop to store and process their most valuable data, the risk of a potential breach of those systems increases exponentially. This practical book not only shows Hadoop administrators and security architects how to protect Hadoop data from unauthorized access, it also shows how to limit the ability of an attacker to corrupt or modify data in the event of a security breach. Authors Ben Spivey and Joey Echeverria provide in-depth information about the security features available in Hadoop, and organize them according to common computer security concepts. You'll also get real-world examples that demonstrate how you can apply these concepts to your use cases. Understand the challenges of securing distributed systems, particularly Hadoop Use best practices for preparing Hadoop cluster hardware as securely as possible Get an overview of the Kerberos network authentication protocol Delve into authorization and accounting principles as they apply to Hadoop Learn how to use mechanisms to protect data in a Hadoop cluster, both in transit and at rest Integrate Hadoop data ingest into enterprise-wide security architecture Ensure that security architecture reaches all the way to end-user access

This book presents task-scheduling techniques for emerging complex parallel architectures including heterogeneous multi-core architectures, warehouse-scale datacenters, and distributed big data processing systems. The demand for high computational capacity has led to the growing popularity of multicore processors, which have become the mainstream in both the research and real-world settings. Yet to date, there is no book exploring the current task-scheduling techniques for the emerging complex parallel architectures. Addressing this gap, the book discusses state-of-the-art task-scheduling techniques that are optimized for different architectures, and which can be directly applied in real parallel systems. Further, the book provides an overview of the latest advances in task-scheduling policies in parallel architectures, and will help readers understand and overcome current and emerging issues in this field.

Hadoop 2 Quick-Start Guide

Moving Beyond MapReduce and Batch Processing with Apache Hadoop 2

19th and 20th International Workshops, JSSPP 2015, Hyderabad, India, May 26, 2015 and JSSPP 2016, Chicago, IL, USA, May 27, 2016, Revised Selected Papers

Learn about big data processing and analytics

Hadoop For Dummies

Large Scale Machine Learning with Python

Sams Teach Yourself Big Data Analytics with Microsoft HDInsight in 24 Hours In just 24 lessons of one hour or less, Sams Teach Yourself Big Data Analytics with Microsoft HDInsight in 24 Hours helps you leverage Hadoop's power on a flexible, scalable cloud platform using Microsoft's newest business intelligence, visualization, and productivity tools. This book's straightforward, step-by-step approach shows you how to provision, configure, monitor, and troubleshoot HDInsight and use Hadoop cloud services to solve real analytics problems. You'll gain more of Hadoop's benefits, with less complexity – even if you're completely new to Big Data analytics. Every lesson builds on what you've already learned, giving you a rock-solid foundation for real-world success. Practical, hands-on examples show you how to apply what you learn Quizzes and exercises help you test your knowledge and stretch your skills Notes and tips point out shortcuts and solutions Learn how to··· · Master core Big Data and NoSQL concepts, value propositions, and use cases · Work with key Hadoop features, such as HDFS2 and YARN · Quickly install, configure, and monitor Hadoop (HDInsight) clusters in the cloud · Automate provisioning, customize clusters, install additional Hadoop projects, and administer clusters · Integrate, analyze, and report with Microsoft BI and Power BI · Automate workflows for data transformation, integration, and other tasks · Use Apache HBase on HDInsight · Use Sqoop or SSIS to move data to or from HDInsight · Perform R-based statistical computing on HDInsight datasets · Accelerate analytics with Apache Spark · Run real-time analytics on high-velocity data streams · Write MapReduce, Hive, and Pig programs Register your book at informit.com/register for convenient access to downloads, updates, and corrections as they become available.

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.

Distributed systems intertwine with our everyday lives. The benefits and current shortcomings of the underpinning technologies are experienced by a wide range of people and their smart devices. With the rise of large-scale IoT and similar distributed systems, cloud bursting technologies, and partial outsourcing solutions, private entities are encouraged to increase their efficiency and offer unparalleled availability and reliability to their users. The Research Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing is a vital reference source that provides valuable insight into current and emergent research occurring within the field of distributed computing. It also presents architectures and service frameworks to achieve highly integrated distributed systems and solutions to integration and efficient management challenges faced by current and future distributed systems. Highlighting a range of topics such as data sharing, wireless sensor networks, and

scalability, this multi-volume book is ideally designed for system administrators, integrators, designers, developers, researchers, academicians, and students.
The digital age has presented an exponential growth in the amount of data available to individuals looking to draw conclusions based on given or collected information across industries. Challenges associated with the analysis, security, sharing, storage, and visualization of large and complex data sets continue to plague data scientists and analysts alike as traditional data processing applications struggle to adequately manage big data. The Handbook of Research on Big Data Storage and Visualization Techniques is a critical scholarly resource that explores big data analytics and technologies and their role in developing a broad understanding of issues pertaining to the use of big data in multidisciplinary fields. Featuring coverage on a broad range of topics, such as architecture patterns, programing systems, and computational energy, this publication is geared towards professionals, researchers, and students seeking current research and application topics on the subject.
Real-Time Analytics
Big Scientific Data Benchmarks, Architecture, and Systems
Hadoop: Data Processing and Modelling
Spark
Hadoop: The Definitive Guide
Storage and Analysis at Internet Scale
The two volumes of this book collect high-quality peer-reviewed research papers presented in the International Conference on ICT for Sustainable Development (ICT4SD 2015) held at Ahmedabad, India during 3 – 4 July 2015. The book discusses all areas of Information and Communication Technologies and its applications in field for engineering and management. The main focus of the volumes are on applications of ICT for Infrastructure, e-Governance, and contemporary technologies advancements on Data Mining, Security, Computer Graphics, etc. The objective of this International Conference is to provide an opportunity for the researchers, academicians, industry persons and students to interact and exchange ideas, experience and expertise in the current trend and strategies for Information and Communication Technologies.
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.
Quickly find solutions to common programming problems encountered while processing big data. Content is presented in the popular problem-solution format. Look up the programming problem that you want to solve. Read the solution. Apply the solution directly in your own code. Problem solved! PySpark Recipes covers Hadoop and its shortcomings. The architecture of Spark, PySpark, and RDD are presented. You will learn to apply RDD to solve day-to-day big data problems. Python and NumPy are included and make it easy for new learners of PySpark to understand and adopt the model. What You Will Learn Understand the advanced features of PySpark2 and SparkSQL Optimize your code Program SparkSQL with Python Use Spark Streaming and Spark MLlib with Python Perform graph analysis with GraphFrames Who This Book Is For Data analysts, Python programmers, big data enthusiasts
If you want to discover one of the latest tools designed to produce stunning Big Data insights, this book features everything you need to get to grips with your data. Whether you are a data architect, developer, or a business strategist, HDInsight adds value in everything from development, administration, and reporting.
A Classroom Approach
PySpark SQL Recipes
Techniques to Analyze and Visualize Streaming Data
Hadoop 2.x Administration Cookbook
Practical Graph Analytics with Apache Giraph
Hadoop Security
Moving beyond MapReduce - learn resource management and big data processing using YARN About This Book Deep dive into YARN components, schedulers, life cycle management and security architecture Create your own Hadoop-YARN applications and integrate big data technologies with YARN Step-by-step guide to provision, manage, and monitor Hadoop-YARN clusters with ease Who This Book Is For This book is intended for those who want to understand what YARN is and how to efficiently use it for the resource management of large clusters. For cluster administrators, this book gives a detailed explanation of provisioning and managing YARN clusters. If you are a Java developer or an open source contributor, this book will help you to drill down the YARN architecture, write your own YARN applications and understand the application

execution phases. This book will also help big data engineers explore YARN integration with real-time analytics technologies such as Spark and Storm. What You Will Learn Explore YARN features and offerings Manage big data clusters efficiently using the YARN framework Create single as well as multi-node Hadoop-YARN clusters on Linux machines Understand YARN components and their administration Gain insights into application execution flow over a YARN cluster Write your own distributed application and execute it over YARN cluster Work with schedulers and queues for efficient scheduling of applications Integrate big data projects like Spark and Storm with YARN In Detail Today enterprises generate huge volumes of data. In order to provide effective services and to make smarter and more intelligent decisions from these huge volumes of data, enterprises use big-data analytics. In recent years, Hadoop has been used for massive data storage and efficient distributed processing of data. The Yet Another Resource Negotiator (YARN) framework solves the design problems related to resource management faced by the Hadoop 1.x framework by providing a more scalable, efficient, flexible, and highly available resource management framework for distributed data processing. This book starts with an overview of the YARN features and explains how YARN provides a business solution for growing big data needs. You will learn to provision and manage single, as well as multi-node, Hadoop-YARN clusters in the easiest way. You will walk through the YARN administration, life cycle management, application execution, REST APIs, schedulers, security framework and so on. You will gain insights about the YARN components and features such as ResourceManager, NodeManager, ApplicationMaster, Container, Timeline Server, High Availability, Resource Localisation and so on. The book explains Hadoop-YARN commands and the configurations of components and explores topics such as High Availability, Resource Localization and Log aggregation. You will then be ready to develop your own ApplicationMaster and execute it over a Hadoop-YARN cluster. Towards the end of the book, you will learn about the security architecture and integration of YARN with big data technologies like Spark and Storm. This book promises conceptual as well as practical knowledge of resource management using YARN. Style and approach Starting with the basics and covering the core concepts with the practical usage, this tutorial is a complete guide to learn and explore YARN offerings.

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.

Bigdata is one of the most demanding markets in the IT sector. If you are an administrator or a have a passion for knowing the internal configurations of Hadoop, then this book is for you. This book enables a professional to learn about Hadoop in terms of installation, configuration, and management. This book will help the reader to jumpstart with Hadoop frameworks, its eco-system components and slowly progress towards learning the administration part of Hadoop. The level of this book goes from beginner to intermediate with 70% hands-on exercises. Some of the techniques that you will learn include, • Installation and configuration of Hadoop cluster • Performing Hadoop Cluster Upgrade • Understanding and implementing HDFS Federation • Understanding and Implementing High Availability • Implementing HA on a Federated Cluster • Zookeeper CLI • Apache Hive Installation and Security • HBase Multi-master setup • Oozie installation, configuration and job submission • Setting up HDFS Quotas • Setting up HDFS NFS gateway • Understanding and implementing rolling upgrade and much more.

This book constitutes the refereed proceedings of the First Workshop on Big Scientific Data Benchmarks, Architecture, and Systems, SDBA 2018, held in Beijing, China, in June 2018. The 10 revised full papers presented were carefully reviewed and selected from 22 submissions. The papers are organized in topical sections on benchmarking; performance optimization; algorithms; big science data framework.

Big Data

Learning YARN

Big Data Management and Processing

YARN Essentials

First Workshop, SDBA 2018, Beijing, China, June 12, 2018, Revised Selected Papers

Apache Hadoop 3 Quick Start Guide

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

Unlock the power of your data with Hadoop 2.X ecosystem and its data warehousing techniques across large data sets About This Book Conquer the mountain of data using Hadoop 2.X tools The authors succeed in creating a context for Hadoop and its ecosystem Hands-on examples and recipes giving the bigger picture and helping you to master Hadoop 2.X data processing platforms Overcome the challenging data processing problems using this exhaustive course with Hadoop 2.X Who This Book Is For This course is for Java developers, who know scripting, wanting a career shift to Hadoop - Big Data segment of the IT industry. So if you are a novice in Hadoop or an expert, this book will make you reach the most advanced level in Hadoop 2.X. What You Will Learn Best practices for setup and configuration of Hadoop clusters, tailoring the system to the problem at hand Integration with relational databases, using Hive for SQL queries and Sqoop for data transfer Installing and maintaining Hadoop 2.X cluster and its ecosystem Advanced Data Analysis using the Hive, Pig, and Map Reduce programs Machine learning principles with libraries such as Mahout and Batch and Stream data processing using Apache Spark Understand the changes involved in the process in the move from Hadoop 1.0 to Hadoop 2.0 Dive into YARN and Storm and use YARN to integrate Storm with Hadoop Deploy Hadoop on Amazon Elastic MapReduce and Discover HDFS replacements and learn about HDFS Federation In Detail As Marc Andreessen has said “ Data is eating the world, ” which can be witnessed today being the age of Big Data, businesses are producing data in huge volumes every day and this rise in tide of data need to be organized and analyzed in a more secured way. With proper and effective use of Hadoop, you can build new-improved models, and based on that you will be able to make the right decisions. The first module, Hadoop beginners Guide will walk you through on understanding Hadoop with very detailed instructions and how to go about using it. Commands are explained using sections called “ What just happened ” for more clarity and understanding. The second module, Hadoop Real World Solutions Cookbook, 2nd edition, is an essential tutorial to effectively implement a big data warehouse in your business, where you get detailed practices on the latest technologies such as YARN and Spark. Big data has become a key basis of competition and the new waves of productivity growth. Hence, once you get familiar with the basics and implement the end-to-end big data use cases, you will start exploring the third module, Mastering Hadoop. So, now the question is if you need to broaden your Hadoop skill set to the next level after you nail the basics and the advance concepts, then this course is indispensable. When you finish this course, you will be able to tackle the real-world scenarios and become a big data expert using the tools and the knowledge based on the various step-by-step tutorials and recipes. Style and approach This course has covered everything right from the basic concepts of Hadoop till you master the advance mechanisms to become a big data expert. The goal here is to help you learn the basic essentials using the step-by-step tutorials and from there moving toward the recipes with various real-world solutions for you. It covers all the important aspects of Hadoop from system designing and configuring Hadoop, machine learning principles with various libraries with chapters illustrated with code fragments and schematic diagrams. This is a compendious course to explore Hadoop from the basics to the most advanced techniques available in Hadoop 2.X.

Over 100 practical recipes to help you become an expert Hadoop administrator About This Book Become an expert Hadoop administrator and perform tasks to optimize your Hadoop Cluster Import and export data into Hive and use Oozie to manage workflow. Practical recipes will help you plan and secure your Hadoop cluster, and make it highly available Who This Book Is For If you are a system administrator with a basic understanding of Hadoop and you want to get into Hadoop administration, this book is for you. It's also ideal if you are a Hadoop administrator who wants a quick reference guide to all the Hadoop administration-related tasks and solutions to commonly occurring problems What You Will Learn Set up the Hadoop architecture to run a Hadoop cluster smoothly Maintain a Hadoop cluster on HDFS, YARN, and MapReduce Understand high availability with Zookeeper and Journal Node Configure Flume for data ingestion and Oozie to run various workflows Tune the Hadoop cluster for optimal performance Schedule jobs on a Hadoop cluster using the Fair and Capacity scheduler Secure your cluster and troubleshoot it for various common pain points In Detail Hadoop enables the distributed storage and processing of large datasets across clusters of computers. Learning how to administer Hadoop is crucial to exploit its unique features. With this book, you will be able to overcome common problems encountered in Hadoop administration. The book begins with laying the foundation by showing you the steps needed to set up a Hadoop cluster and its various nodes. You will get a better understanding of how to maintain Hadoop cluster, especially on the HDFS layer and using YARN and MapReduce. Further on, you will explore durability and high availability of a Hadoop cluster. You'll get a better understanding of the schedulers in Hadoop and how to configure and use them for your tasks. You will also get hands-on experience with the backup and recovery options and the performance tuning aspects of Hadoop. Finally, you will get a better understanding of troubleshooting, diagnostics, and best practices in Hadoop administration. By the end of this book, you will have a proper understanding of working with Hadoop clusters and will also be able to secure, encrypt it, and configure auditing for your Hadoop clusters. Style and approach This book contains short recipes that will help you run a Hadoop cluster efficiently. The recipes are solutions to real-life problems that administrators encounter while working with a Hadoop cluster

Over 70 recipes to help you use Apache Spark as your single big data computing platform and master its libraries About This Book This book contains recipes on how to use Apache Spark as a unified compute engine Cover how to connect various source systems to Apache Spark Covers various parts of machine learning including supervised/unsupervised learning & recommendation engines Who This Book Is For This book is for data engineers, data scientists, and those who want to implement Spark for real-time data processing. Anyone who is using Spark (or is planning to) will benefit from this book. The book assumes you have a basic knowledge of Scala as a programming language. What You Will Learn Install and configure Apache Spark with various cluster managers & on AWS Set up a development environment for Apache Spark including Databricks Cloud notebook Find out how to operate on data in Spark with schemas Get to grips with real-time streaming analytics using Spark Streaming & Structured Streaming Master supervised learning and unsupervised learning using MLlib Build a recommendation engine using MLlib Graph processing using GraphX and GraphFrames libraries Develop a set of common applications or project types, and solutions that solve complex big data problems In Detail While Apache Spark 1.x gained a lot of traction and adoption in the early years, Spark 2.x delivers notable improvements in the areas of API, schema awareness, Performance, Structured Streaming, and simplifying building blocks to build better, faster, smarter, and more accessible big data applications. This book uncovers all these features in the form of structured recipes to analyze and mature large and complex sets of data. Starting with installing and configuring Apache Spark with various cluster managers, you will learn to set up development environments. Further on, you will be introduced to working with RDDs, DataFrames and Datasets to operate on schema aware data, and real-time streaming with various sources such as Twitter Stream and Apache Kafka. You will also work through recipes on machine learning, including supervised learning, unsupervised learning & recommendation engines in Spark. Last but not least, the final few chapters delve deeper into the concepts of graph processing using GraphX, securing your implementations, cluster optimization, and troubleshooting. Style and approach This book is packed with intuitive recipes supported with line-by-line explanations to help you understand Spark 2.x's real-time processing capabilities and deploy scalable big data solutions. This is a valuable resource for data scientists and those working on large-scale data projects.

Learning from Imbalanced Data Sets
Learning Hadoop 2
Transactions on Large-Scale Data- and Knowledge-Centered Systems XLV
Principles and Paradigms
Big Data Cluster Computing in Production
PySpark Recipes

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

From the Foreword: "Big Data Management and Processing is [a] state-of-the-art book that deals with a wide range of topical themes in the field of Big Data. The book, which probes many issues related to this exciting and rapidly growing field, covers processing, management, analytics, and applications... [It] is a very valuable addition to the literature. It will serve as a source of up-to-date research in this continuously developing area. The book also provides an opportunity for researchers to explore the use of advanced computing technologies and their impact on enhancing our capabilities to conduct more sophisticated studies." ---Sartaj Sahni, University of Florida, USA "Big Data Management and Processing covers the latest Big Data research results in processing, analytics, management and applications. Both fundamental insights and representative applications are provided. This book is a timely and valuable resource for students, researchers and seasoned practitioners in Big Data fields. --Hai Jin, Huazhong University of Science and Technology, China Big Data Management and Processing explores a range of big data related issues and their impact on the design of new computing systems. The twenty-one chapters were carefully selected and feature contributions from several outstanding researchers. The book endeavors to strike a balance between theoretical and practical coverage of innovative problem solving techniques for a range of platforms. It serves as a repository of paradigms, technologies, and applications that target different facets of big data computing systems. The first part of the book explores energy and resource management issues, as well as legal compliance and quality management for Big Data. It covers In-Memory computing and In-Memory data grids, as well as co-scheduling for high performance computing applications. The second part of the book includes comprehensive coverage of Hadoop and Spark, along with security, privacy, and trust challenges and solutions. The latter part of the book covers mining and clustering in Big Data, and includes applications in genomics, hospital big data processing, and vehicular cloud computing. The book also analyzes funding for Big Data projects.

Get ready to unlock the power of your data. With the fourth edition of this comprehensive guide, you â ??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 â ??ll learn about recent changes to Hadoop, and explore new case studies on Hadoop â ??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

This unique text helps make sense of big data in engineering applications using tools and techniques from signal processing. It presents fundamental signal

processing theories and software implementations, reviews current research trends and challenges, and describes the techniques used for analysis, design and optimization. Readers will learn about key theoretical issues such as data modelling and representation, scalable and low-complexity information processing and optimization, tensor and sublinear algorithms, and deep learning and software architecture, and their application to a wide range of engineering scenarios. Applications discussed in detail include wireless networking, smart grid systems, and sensor networks and cloud computing. This is the ideal text for researchers and practising engineers wanting to solve practical problems involving large amounts of data, and for students looking to grasp the fundamentals of big data analytics.

Protecting Your Big Data Platform

Big Data Analytics with Microsoft HDInsight in 24 Hours, Sams Teach Yourself

Proceedings of International Conference on ICT for Sustainable Development

With HiveQL, Dataframe and Graphframes

Volume 2

Hadoop in Practice

This multi-contributed handbook focuses on the latest workings of IoT (internet of Things) and Big Data. As the resources are limited, it's the endeavor of the authors to support and bring the information into one resource. The book is divided into 4 sections that covers IoT and technologies, the future of Big Data, algorithms, and case studies showing IoT and Big Data in various fields such as health care, manufacturing and automation. Features Focuses on the latest workings of IoT and Big Data Discusses the emerging role of technologies and the fast-growing market of Big Data Covers the movement toward automation with hardware, software, and sensors, and trying to save on energy resources Offers the latest technology on IoT Presents the future horizons on Big Data

A fast paced guide that will help you learn about Apache Hadoop 3 and its ecosystem Key Features Set up, configure and get started with Hadoop to get useful insights from large data sets Work with the different components of Hadoop such as MapReduce, HDFS and YARN Learn about the new features introduced in Hadoop 3 Book Description Apache Hadoop is a widely used distributed data platform. It enables large datasets to be efficiently processed instead of using one large computer to store and process the data. This book will get you started with the Hadoop ecosystem, and introduce you to the main technical topics, including MapReduce, YARN, and HDFS. The book begins with an overview of big data and Apache Hadoop. Then, you will set up a pseudo Hadoop development environment and a multi-node enterprise Hadoop cluster. You will see how the parallel programming paradigm, such as MapReduce, can solve many complex data processing problems. The book also covers the important aspects of the big data software development lifecycle, including quality assurance and control, performance, administration, and monitoring. You will then learn about the Hadoop ecosystem, and tools such as Kafka, Sqoop, Flume, Pig, Hive, and HBase. Finally, you will look at advanced topics, including real time streaming using Apache Storm, and data analytics using Apache Spark. By the end of the book, you will be well versed with different configurations of the Hadoop 3 cluster. What you will learn Store and analyze data at scale using HDFS, MapReduce and YARN Install and configure Hadoop 3 in different modes Use Yarn effectively to run different applications on Hadoop based platform Understand and monitor how Hadoop cluster is managed Consume streaming data using Storm, and then analyze it using Spark Explore Apache Hadoop ecosystem components, such as Flume, Sqoop, HBase, Hive, and Kafka Who this book is for Aspiring Big Data professionals who want to learn the essentials of Hadoop 3 will find this book to be useful. Existing Hadoop users who want to get up to speed with the new features introduced in Hadoop 3 will also benefit from this book. Having knowledge of Java programming will be an added advantage.

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.

Carry out data analysis with PySpark SQL, graphframes, and graph data processing using a problem-solution approach. This book provides solutions to problems related to dataframes, data manipulation summarization, and exploratory analysis. You will improve your skills in graph data analysis using graphframes and see how to optimize your PySpark SQL code. PySpark SQL Recipes starts with recipes on creating dataframes from different types of data source, data aggregation and summarization, and exploratory data analysis using PySpark SQL. You ' ll also discover how to solve problems in graph analysis using graphframes. On completing this book, you ' ll have ready-made code for all your PySpark SQL tasks, including creating dataframes using data from different file formats as well as from SQL or NoSQL databases. What You Will Learn Understand PySpark SQL and its advanced features Use SQL and HiveQL with PySpark SQL Work with structured streaming Optimize PySpark SQL Master graphframes and graph processing Who This Book Is ForData scientists, Python programmers, and SQL programmers.

Handbook of Research on Big Data Storage and Visualization Techniques
Beginning Apache Hadoop Administration
ICT4SD 2015 Volume 1
Special Issue on Data Management and Knowledge Extraction in Digital Ecosystems
Task Scheduling for Multi-core and Parallel Architectures
Signal Processing and Networking for Big Data Applications

"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

This book constitutes the thoroughly refereed post-conference proceedings of the 19th and 20th International Workshop on Job Scheduling Strategies for Parallel Processing, JSSPP 2015 and 2016, held respectively in Hyderabad, India, on May 26, 2015 and in Chicago, IL, USA, on May 27, 2016. The 14 revised full papers presented (7 papers in 2015 and 7 papers in 2016) were carefully reviewed and selected from 28 submissions (14 in 2015 and 14 in 2016). The papers cover the following topics: parallel scheduling raising challenges multiple levels of abstractions; node level parallelism; minimization of energy consumption in task migration within a many-core chip; task replication in real-time scheduling context; data-driven approach to schedule GPU load; the use of lock-free data structures in OS scheduler; the influence between user behaviour (think time, more precisely) and parallel scheduling; Evalix, a predictor for job resource consumption; sophisticated and realistic simulation; space-filling curves leading to better scheduling of large-scale computers; discussion of real-life production experiences.

The LNCS journal Transactions on Large-Scale Data- and Knowledge-Centered Systems focuses on data management, knowledge discovery, and knowledge processing, which are core and hot topics in computer science. Since the 1990s, the Internet has become the main driving force behind application development in all domains. An increase in the demand for resource sharing (e.g., computing resources, services, metadata, data sources) across different sites connected through networks has led to an evolution of data- and knowledge-management systems from centralized systems to decentralized systems enabling large-scale distributed applications providing high scalability. This, the 45th issue of Transactions on Large-Scale Data- and Knowledge-Centered Systems, contains eight revised selected regular papers. Topics covered include data analysis, information extraction, blockchains, and big data.

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

A Problem-Solution Approach with PySpark2
HDInsight Essentials - Second Edition
Big Data Concepts, Theories, and Applications
Apache Hadoop YARN

Algorithms and Architectures for Parallel Processing