Hadoop 2 Quick Start Guide Learn The Essentials Of Big Data Computing In The Apache Hadoop 2 Ecosystem Addison Wesley Data Analytics Series

This book includes the outcomes of the International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD-2018), held in Tangier, Morocco on July 12–14, 2018. Presenting the latest research in the field of computing sciences and information technology, it discusses new challenges and provides valuable insights into the field, the goal being to stimulate debate, and to promote closer interaction and interdisciplinary collaboration between researchers and practitioners. Though chiefly intended for researchers and practitioners in advanced information technology management and networking, the book will also be of interest to those engaged in emerging fields such as data science and analytics, big data, internet of things, smart networked systems, artificial intelligence, expert systems and cloud computing.

See a Mesos-based big data stack created and the components used. You will use currently available Apache full and incubating systems. The components are introduced by example and you learn how they work together. In the Complete Guide to Open Source Big Data Stack, the author begins by creating a private cloud and then installs and examines Apache Brooklyn. After that, he uses each chapter to introduce one piece of the big data stack—sharing how to source the software and how to install it. You learn by simple example, step by step and chapter by chapter, as a real big data stack is created. The book concentrates on Apache-based systems and shares detailed examples of cloud storage, release management, resource management, processing, queuing, frameworks, data visualization, and more. What You'll Learn Install a private cloud onto the local cluster using Apache cloud stack Source, install, and configure Apache: Brooklyn, Mesos, Kafka, and Zeppelin See how Brooklyn can be used to install Mule ESB on a cluster and Cassandra in the cloud Install and use DCOS for big data processing Use Apache Spark for big data stack data processing Who This Book Is For Developers, architects, IT project managers, database administrators, and others charged with developing or supporting a big data system. It is also for anyone interested in Hadoop or big data, and those experiencing problems with data size.

Learn about the fastest-growing open source project in the world, and find out how it revolutionizes big data analytics About This Book Exclusive guide that covers how to get up and running with fast data processing using Apache Spark Explore and exploit various possibilities with Apache Spark using real-world use cases in this book Want to perform efficient data processing at real time? This book will be your one-stop solution. Who This Book Is For This guide appeals to big data engineers, analysts, architects, software engineers, even technical managers who need to perform efficient data processing on Hadoop at real time. Basic familiarity with Java or Scala will be helpful. The assumption is that readers will be from a mixed background, but would be typically people with background in engineering/data science with no prior Spark experience and want to understand how Spark can help them on their analytics journey. What You Will Learn Get an overview of big data analytics and its importance for organizations and data professionals Delve into Spark to see how it is different from existing processing platforms Understand the intricacies of various file formats, and how to process them with Apache Spark. Realize how to deploy Spark with YARN, MESOS or a Stand-alone cluster manager. Learn the concepts of Spark SQL, SchemaRDD, Caching and working with Hive and Parquet file formats Understand the architecture of Spark MLLib while discussing some of the off-the-shelf algorithms that come with Spark. Introduce yourself to the deployment and usage of SparkR. Walk through the importance of Graph computation and the graph processing systems available in the market Check the real world example of Spark by building a recommendation engine with Spark using ALS. Use a Telco data set, to predict customer churn using Random Forests. In Detail Spark juggernaut keeps on rolling and getting more and more momentum each day. Spark provides key capabilities in the form of Spark SQL, Spark Streaming, Spark ML and Graph X all accessible via Java, Scala, Python and R. Deploying the key capabilities is crucial whether it is on a Standalone framework or as a part of existing Hadoop installation and configuring with Yarn and Mesos. The next part of the journey after installation is using key components, APIs, Clustering, machine learning APIs, data pipelines, parallel programming. It is important to understand why each framework component is key, how widely it is being used, its stability and pertinent use cases. Once we understand the individual components, we will take a couple of real life advanced analytics examples such as 'Building a Recommendation system', 'Predicting customer churn' and so on. The objective of these real life examples is to give the reader confidence of using Spark for real-world problems. Style and approach With the help of practical examples and real-world use cases, this guide will take you from scratch to building efficient data applications using Apache Spark. You will learn all about this excellent data processing engine in a step-by-step manner, taking one aspect of it at a time. This highly practical guide will include how to work with data pipelines, dataframes, clustering, SparkSQL, parallel programming, and such insightful topics with the help of real-world use cases.

As technology weaves itself more tightly into everyday life, socio-economic development has become intricately tied to these ever-evolving innovations. Technology management is now an integral element of sound business practices, and this revolution has opened up many opportunities for global communication. However, such swift change warrants greater research that can foresee and possibly prevent future complications within and between organizations. The Handbook of Research on Engineering Innovations and Technology Management in Organizations is a collection of innovative research that explores global concerns in the applications of technology to business and the explosive growth that resulted. Highlighting a wide range of topics such as cyber security, legal practice, and artificial intelligence, this book is ideally designed for engineers, manufacturers, technology managers, technology developers, IT specialists, productivity consultants, executives, lawyers, programmers, managers, policymakers, academicians, researchers, and students.

???????????

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

If you're looking for a scalable storage solution to accommodate a virtually endless amount of data, this book shows you how Apache HBase can fulfill your needs. As the open source implementation of Google's BigTable architecture, HBase scales to billions of rows and millions of columns, while ensuring that write and read performance remain constant. Many IT executives are asking pointed questions

about HBase. This book provides meaningful answers, whether you're evaluating this non-relational database or planning to put it into practice right away. Discover how tight integration with Hadoop makes scalability with HBase easier Distribute large datasets across an inexpensive cluster of commodity servers Access HBase with native Java clients, or with gateway servers providing REST, Avro, or Thrift APIs Get details on HBase's architecture, including the storage format, write-ahead log, background processes, and more Integrate HBase with Hadoop's MapReduce framework for massively parallelized data processing jobs Learn how to tune clusters, design schemas, copy tables, import bulk data, decommission nodes, and many other tasks

??????????????????, ??????????????, ??????????????????????????????.

Combine advanced analytics including Machine Learning, Deep Learning Neural Networks and Natural Language Processing with modern scalable technologies including Apache Spark to derive actionable insights from Big Data in real-time Key Features Make a hands-on start in the fields of Big Data, Distributed Technologies and Machine Learning Learn how to design, develop and interpret the results of common Machine Learning algorithms Uncover hidden patterns in your data in order to derive real actionable insights and business value Book Description Every person and every organization in the world manages data, whether they realize it or not. Data is used to describe the world around us and can be used for almost any purpose, from analyzing consumer habits to fighting disease and serious organized crime. Ultimately, we manage data in order to derive value from it, and many organizations around the world have traditionally invested in technology to help process their data faster and more efficiently. But we now live in an interconnected world driven by mass data creation and consumption where data is no longer rows and columns restricted to a spreadsheet, but an organic and evolving asset in its own right. With this realization comes major challenges for organizations: how do we manage the sheer size of data being created every second (think not only spreadsheets and databases, but also social media posts, images, videos, music, blogs and so on)? And once we can manage all of this data, how do we derive real value from it? The focus of Machine Learning with Apache Spark is to help us answer these questions in a hands-on manner. We introduce the latest scalable technologies to help us manage and process big data. We then introduce advanced analytical algorithms applied to real-world use cases in order to uncover patterns, derive actionable insights, and learn from this big data. What you will learn Understand how Spark fits in the context of the big data ecosystem Understand how to deploy and configure a local development environment using Apache Spark Understand how to design supervised and unsupervised learning models Build models to perform NLP, deep learning, and cognitive services using Spark ML libraries Design real-time machine learning pipelines in Apache Spark Become familiar with advanced techniques for processing a large volume of data by applying machine learning algorithms Who this book is for This book is aimed at Business Analysts, Data Analysts and Data Scientists who wish to make a hands-on start in order to take advantage of modern Big Data technologies combined with Advanced Analytics.

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.

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

Hadoop 2 Quick-Start GuideLearn the Essentials of Big Data Computing in the Apache Hadoop 2 EcosystemAddison-Wesley Professional

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
Get ready to unlock the power of your data. With the fourth edition of this comprehensive guide, you'll learn how to build and maintain reliable, scalable, distributed systems with Apache Hadoop. This book is ideal for programmers looking to analyze datasets of any size, and for administrators who want to set up and run Hadoop clusters. Using Hadoop 2 exclusively, author Tom White presents new chapters on YARN and several Hadoop-related projects such as Parquet, Flume, Crunch, and Spark. You'll learn about recent changes to Hadoop, and explore new case studies on Hadoop's role in healthcare systems and genomics data processing. Learn fundamental components such as MapReduce, HDFS, and YARN Explore MapReduce in depth, including steps for developing applications with it Set up and maintain a Hadoop cluster running HDFS and MapReduce on YARN Learn two data formats: Avro for data serialization and Parquet for nested data Use data ingestion tools such as Flume (for streaming data) and Sqoop (for bulk data transfer) Understand how high-level data processing tools like Pig, Hive, Crunch, and Spark work with Hadoop Learn the HBase distributed database and the ZooKeeper distributed configuration service

If you are a Hadoop programmer who wants to learn about Flume to be able to move datasets into Hadoop in a timely and replicable manner, then this book is ideal for you. No prior knowledge about Apache Flume is necessary, but a basic knowledge of Hadoop and the Hadoop File System (HDFS) is assumed.

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

??????????????? ??????????????? ????????????????? ??? ??????? ??? ?Wired????????? ??? ???? ?????????????????????????? ???????????? —— ????Lawrence Lessig????????????????? ?????????????????????????? ?????????????????????????? ?????????????????????? ?????????????????????????? ?? ????????????????????????????????? ???????????????????????????? ?????????????????? ??????????????????????????? ???????????????…… ???????????????????????????? ?????????????????????????? ?????????????????????????????? ??????????????????? ?????????????????????????

This textbook adopts a unique approach to helping developers and CS students learn Hadoop MapReduce programming fast in an easy-to-setup, virtual 4-node Linux YARN cluster on a Windows laptop. Rather than filled with disjointed, piecemeal code snippets to show Hadoop MapReduce programming features one at a time, it is designed to place your total Hadoop MapReduce programming learning process in a common application context of mining customer spending patterns ensconced in large volumes of credit card transaction record data. Precise, end-to-end procedures are given to help you set up your Hadoop MapReduce development environment quickly on Eclipse with Maven on Windows. Step-by-step procedures are also given on how to set up a four-node Linux cluster at minimum so that you can run your MapReduce programs not only in local but also in standalone and fully distributed mode on a real cluster. In fact, all MapReduce programs presented in the book have been tested and verified on such a Linux cluster. This textbook mainly focuses on teaching Hadoop MapReduce programming in a scientific, objective, quantitative approach. Rather than heavily relying on subjective, verbose (and sometimes even pompous) textual descriptions with sparse code snippets, this textbook uses Hadoop Java APIs, Hadoop configuration parameters, complete MapReduce programs and their execution logs and outputs to demonstrate how Hadoop MapReduce framework works and how to write MapReduce programs. Specifically, this text covers the following subjects: * Introduction to Hadoop * Setting up a Linux Hadoop Cluster * The Hadoop Distributed FileSystem * MapReduce Job Orchestration and Workflows * Basic MapReduce Programming * Advanced MapReduce Programming * Hadoop Streaming * Hadoop Administration No matter what role you play on your team, this text can help you gain truly applicable Hadoop skills in a most effective and efficient manner. The book can also be used as a supplementary textbook for a distributed computing or Hadoop course offered to upper-division CS students.

???PMBOK??(?5?)?????,???PMBOK??(?5?)????,???47?????????????????????????,?????.?????????????,??????,??????,???????.

"This book is a critically needed resource for the newly released Apache Hadoop 2.0, highlighting YARN as the significant breakthrough that broadens Hadoop beyond the MapReduce paradigm." —From the Foreword by Raymie Stata, CEO of Altiscale The Insider's Guide to Building Distributed, Big Data Applications with Apache Hadoop™ YARN 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 Hadoop™ YARN, two Hadoop technical leaders show you how to develop new applications and adapt existing code to fully leverage these revolutionary advances. YARN project founder Arun Murthy and project lead Vinod Kumar Vavilapalli demonstrate how YARN increases scalability and cluster utilization, enables new programming models and services, and opens new options beyond Java and batch processing. They walk you through the entire YARN project lifecycle, from installation through deployment. You'll find many examples drawn from the authors' cutting-edge experience—first as Hadoop's earliest developers and implementers at Yahoo! and now as Hortonworks developers moving the platform forward and helping customers succeed with it. Coverage includes YARN's goals, design, architecture, and components—how it expands the Apache Hadoop ecosystem Exploring YARN on a single node Administering YARN clusters and Capacity Scheduler Running existing MapReduce applications Developing a large-scale clustered YARN application Discovering new open source frameworks that run under YARN

Data is arriving faster than you can process it and the overall volumes keep growing at a rate that keeps you awake at night. Hadoop can help you tame the data beast. Effective use of Hadoop however

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

Apache Spark is a flexible in-memory framework that allows processing of both batch and real-time data. Its unified engine has made it quite popular for big data use cases. This book will help you to quickly get started with Apache Spark 2.0 and write efficient big data applications for a variety of use cases.

The book Intelligent Systems and Applications - Proceedings of the 2020 Intelligent Systems Conference is a remarkable collection of chapters covering a wider range of topics in areas of intelligent systems and artificial intelligence and their applications to the real world. The Conference attracted a total of 545 submissions from many academic pioneering researchers, scientists, industrial engineers, students from all around the world. These submissions underwent a double-blind peer review process. Of those 545 submissions, 177 submissions have been selected to be included in these proceedings. As intelligent systems continue to replace and sometimes outperform human intelligence in decision-making processes, they have enabled a larger number of problems to be tackled more effectively.This branching out of computational intelligence in several directions and use of intelligent systems in everyday applications have created the need for such an international conference which serves as a venue to report on up-to-the-minute innovations and developments. This book collects both theory and application based chapters on all aspects of artificial intelligence, from classical to intelligent scope. We hope that readers find the volume interesting and valuable; it provides the state of the art intelligent methods and techniques for solving real world problems along with a vision of the future research.

??????????????,????,?????,???,??,????,????,??/??,??,??????,?????,??????MINIX 3????.

Due to the increasing availability of affordable internet services, the number of users, and the need for a wider range of multimedia-based applications, internet usage is on the rise. With so many users and such a large amount of data, the requirements of analyzing large data sets leads to the need for further advancements to information processing. Big Data Processing With Hadoop is an essential reference source that discusses possible solutions for millions of users working with a variety of data applications, who expect fast turnaround responses, but encounter issues with processing data at the rate it comes in. Featuring research on topics such as market basket analytics, scheduler load simulator, and writing YARN applications, this book is ideally designed for IoT professionals, students, and engineers seeking coverage on many of the real-world challenges regarding big data.

Databases Illuminated, Third Edition Includes Navigate 2 Advantage Access combines database theory with a practical approach to database design and implementation. Strong pedagogical features, including accessible language, real-world examples, downloadable code, and engaging hands-on projects and lab exercises create a text with a unique combination of theory and student-oriented activities. Providing an integrated, modern approach to databases, Databases Illuminated, Third Edition is the essential text for students in this expanding field.

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

Get started fast with Apache Hadoop 2, with the first easy, accessible guide to this revolutionary Big Data technology. Building on his unsurpassed experience teaching Hadoop and Big Data, Dr. Douglas Eadline covers all the basics you need to know to install and use Hadoop 2 on both personal computers and servers, and navigate the entire Apache Hadoop ecosystem. Eadline demystifies Hadoop 2, explains the problems it solves, shows how it relates to Big Data, and demonstrates both administrators and users work with it. He explains the central role of MapReduce in Hadoop 1, and how (and why) YARN and Hadoop 2 move beyond MapReduce. You'll find essential information on: Planning and performing Hadoop 2 installations -- including decisions about hardware, software, clustering, and HDFS Using the Hadoop Distributed File System (HDFS) and working around its tradeoffs Running and benchmarking Hadoop 2 programs Working with MapReduce -- including basic programming examples Using higher-level tools, including Pig and Hive Getting started with Apache Hadoop YARN frameworks Administering Hadoop 2 with Ambari, rmadmin, and automated scripts From its Getting Started checklist/flowchart to its roadmap of additional resources, Hadoop 2 Quick-Start Guide is your perfect Hadoop 2 starting point -- and your fastest way to start mastering Big Data.

"Hadoop and Spark Fundamentals LiveLessons provides 9+ hours of video introduction to the Apache Hadoop Big Data ecosystem. The tutorial includes background information and explains the core components of Hadoop, including Hadoop Distributed File Systems (HDFS), MapReduce, the YARN resource manager, and YARN Frameworks. In addition, it demonstrates how to use Hadoop at several levels, including the native Java interface, C++ pipes, and the universal streaming program interface. Examples include how to use benchmarks and high-level tools, including the Apache Pig scripting language, Apache Hive "SQL-like" interface, Apache Flume for streaming input, Apache Sqoop for import and export of relational data, and Apache Oozie for Hadoop workflow management. In addition, there is comprehensive coverage of Spark, PySpark, and the Zeppelin web-GUI. The steps for easily installing a working Hadoop/Spark system

on a desktop/laptop and on a local stand-alone cluster using the powerful Ambari GUI are also included. All software used in these LiveLessons is open source and freely available for your use and experimentation. A bonus lesson includes a quick primer on the Linux command line as used with Hadoop and Spark."--Resource description page.

A guide to cloud computing for students, scientists, and engineers, with advice and many hands-on examples. The emergence of powerful, always-on cloud utilities has transformed how consumers interact with information technology, enabling video streaming, intelligent personal assistants, and the sharing of content. Businesses, too, have benefited from the cloud, outsourcing much of their information technology to cloud services. Science, however, has not fully exploited the advantages of the cloud. Could scientific discovery be accelerated if mundane chores were automated and outsourced to the cloud? Leading computer scientists Ian Foster and Dennis Gannon argue that it can, and in this book offer a guide to cloud computing for students, scientists, and engineers, with advice and many hands-on examples. The book surveys the technology that underpins the cloud, new approaches to technical problems enabled by the cloud, and the concepts required to integrate cloud services into scientific work. It covers managing data in the cloud, and how to program these services; computing in the cloud, from deploying single virtual machines or containers to supporting basic interactive science experiments to gathering clusters of machines to do data analytics; using the cloud as a platform for automating analysis procedures, machine learning, and analyzing streaming data; building your own cloud with open source software; and cloud security. The book is accompanied by a website, Cloud4SciEng.org, that provides a variety of supplementary material, including exercises, lecture slides, and other resources helpful to readers and instructors.

???11??????????????????????1~6??????Hadoop???Hadoop????????????HDFS?????MapReduce?????Yarn?????????????7~11??????????Zookeeper?Hadoop???Hbase?Hadoop????Hive?Hadoop????Flume?OTA?????????

2008????????????????????? 50????????49?????? 35???????????????? 2012??????????????? ???50?????? ??????????????????????????? ???????????????????????????? ???????? ????????????????? ?????????????? ????????????????????? ???????????????????20 ????????1??????????1? ?????????????10? ????????2012?????????1? ?????????????2012????????????? ??????????2012????????????? ?????????????10?????????!? ?2009??????????????????????????????? ?2012?????????????????????????? ??????????? ???????????????????????911????? ?2007?????????????????????????????? ???????? (big data)???! ???(????:????????????????????) ?? ???;?? ?? ???(???????????????: ???8?) ????????????: ???????????????????????? ??? ??????????????????????????? ?????????????????? ??????????????????????????????????????? ??? ??? ??

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.

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