Hadoop Operations

When somebody should go to the ebook stores, search establishment by shop, shelf by shelf, it is in point of fact problematic. This is why we give the ebook compilations in this website. It will certainly ease you to see guide **Hadoop Operations** as you such as.

By searching the title, publisher, or authors of guide you in reality want, you can discover them rapidly. In the house, workplace, or perhaps in your method can be every best place within net connections. If you mean to download and install the Hadoop Operations , it is entirely simple then, previously currently we extend the link to buy and make bargains to download and install Hadoop Operations for that reason simple!

R for Programmers - Dan Zhang 2016-01-06
Unlike other books about R, written from the perspective of statistics, R for Programmers: Mastering the Tools is written from the perspective of programmers, providing a channel for programmers with expertise in other programming languages to quickly understand R. The contents are divided into four sections: The first section consists of the basic

Hadoop: Data Processing and Modelling - Garry Turkington 2016-08-31
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.

Hadoop Security - Ben Spivey 2015-06-29
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
Applied Big Data Analytics in Operations Management - Kumar, Manish 2016-09-30
Operations management is a tool by which companies can effectively meet customers' needs using the least amount of resources necessary. With the emergence of sensors and smart metering, big data is becoming an intrinsic part of modern operations management. Applied Big Data Analytics in Operations Management enumerates the challenges and creative solutions and tools to apply when using big data in

operations management. Outlining revolutionary concepts and applications that help businesses predict customer behavior along with applications of artificial neural networks, predictive analytics, and opinion mining on business management, this comprehensive publication is ideal for IT professionals, software engineers, business professionals, managers, and students of management.

The Stances of e-Government - Puneet Kumar 2018-11-01

This book focuses on the three inevitable facets of e-government, namely policies, processes and technologies. The policies discusses the genesis and revitalization of government policies; processes talks about ongoing e-government practices across developing countries; technology reveals the inclusion of novel technologies.

Hadoop Application Architectures - Mark Grover 2015-06-30

Get expert guidance on architecting end-to-end data management solutions with Apache Hadoop. While many sources explain how to use various components in the Hadoop ecosystem, this practical book takes you through architectural considerations necessary to tie those components together into a complete tailored application, based on your particular use case. To reinforce those lessons, the book's second section provides detailed examples of architectures used in some of the most commonly found Hadoop applications. Whether you're designing a new Hadoop application, or planning to integrate Hadoop into your existing data infrastructure, Hadoop Application Architectures will skillfully guide you through the process. This book covers: Factors to consider when using Hadoop to store and model data Best practices for moving

data in and out of the system Data processing frameworks, including MapReduce, Spark, and Hive Common Hadoop processing patterns, such as removing duplicate records and using windowing analytics Giraph, GraphX, and other tools for large graph processing on Hadoop Using workflow orchestration and scheduling tools such as Apache Oozie Near-real-time stream processing with Apache Storm, Apache Spark Streaming, and Apache Flume Architecture examples for clickstream analysis, fraud detection, and data warehousing

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

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

Sqoop Perform large-scale data processing with the Pig query language Analyze datasets with Hive, Hadoop's data warehousing system Take advantage of HBase for structured and semi-structured data, and ZooKeeper for building distributed systems

Hadoop Operations - Eric Sammer 2012-10-09

For system administrators tasked with the job of maintaining large and complex Hadoop clusters, this book explains the particulars of Hadoop operations, from planning, installing, and configuring the system to providing ongoing maintenance.

MapReduce Design Patterns - Donald Miner 2012-11-21

Until now, design patterns for the MapReduce framework have been scattered among various research papers, blogs, and books. This handy guide brings together a unique collection of valuable MapReduce patterns that will save you time and effort regardless of the domain, language, or development framework you're using. Each pattern is explained in context, with pitfalls and caveats clearly identified to help you avoid common design mistakes when modeling your big data architecture. This book also provides a complete overview of MapReduce that explains its origins and implementations, and why design patterns are so important. All code examples are written for Hadoop. Summarization patterns: get a top-level view by summarizing and grouping data Filtering patterns: view data subsets such as records generated from one user Data organization patterns: reorganize data to work with other systems, or to make MapReduce analysis easier Join patterns: analyze different datasets together to discover interesting relationships Metapatterns: piece together several

patterns to solve multi-stage problems, or to perform several analytics in the same job Input and output patterns: customize the way you use Hadoop to load or store data "A clear exposition of MapReduce programs for common data processing patterns—this book is indespensible for anyone using Hadoop." -- Tom White, author of Hadoop: The Definitive Guide

Moving Hadoop to the Cloud - Bill Havanki 2017-07-14 Until recently, Hadoop deployments existed on hardware owned and run by organizations. Now, of course, you can acquire the computing resources and network connectivity to run Hadoop clusters in the cloud. But there's a lot more to deploying Hadoop to the public cloud than simply renting machines. This hands-on guide shows developers and systems administrators familiar with Hadoop how to install, use, and manage cloud-born clusters efficiently. You'll learn how to architect clusters that work with cloud-provider features—not just to avoid pitfalls, but also to take full advantage of these services. You'll also compare the Amazon, Google, and Microsoft clouds, and learn how to set up clusters in each of them. Learn how Hadoop clusters run in the cloud, the problems they can help you solve, and their potential drawbacks Examine the common concepts of cloud providers, including compute capabilities, networking and security, and storage Build a functional Hadoop cluster on cloud infrastructure, and learn what the major providers require Explore use cases for high availability, relational data with Hive, and complex analytics with Spark Get patterns and practices for running cloud clusters, from designing for price and security to dealing with maintenance

Hadoop For Dummies - Dirk deRoos 2014-03-21 Let Hadoop For Dummies help harness the power of yourdata and rein in the information overload Big data has become big business, and companies and organizationsof all sizes are struggling to find ways to retrieve valuableinformation from their massive data sets with becoming overwhelmed.Enter Hadoop and this easy-to-understand For Dummiesguide. Hadoop For Dummies helps readers understand thevalue of big data, make a business case for using Hadoop, navigatethe Hadoop ecosystem, and build and manage Hadoop applications andclusters. Explains the origins of Hadoop, its economic benefits, and itsfunctionality and practical applications Helps you find your way around the Hadoop ecosystem, programMapReduce, utilize design patterns, and get your Hadoop cluster upand running quickly and easily Details how to use Hadoop applications for data mining, webanalytics and personalization, large-scale text processing, datascience, and problem-solving Shows you how to improve the value of your Hadoop cluster,maximize your investment in Hadoop, and avoid common pitfalls whenbuilding your Hadoop cluster From programmers challenged with building and maintainingaffordable, scaleable data systems to administrators who must dealwith huge volumes of information effectively and efficiently, thishow-to has something to help you with Hadoop.

Apache Sqoop Cookbook - Kathleen Ting 2013-07-02 Integrating data from multiple sources is essential in the age of big data, but it can be a challenging and time-consuming task. This handy cookbook provides dozens of ready-to-use recipes for using Apache

Sqoop, the command-line interface application that optimizes data transfers between relational databases and Hadoop. Sqoop is both powerful and bewildering, but with this cookbook's problem-solution-discussion format, you'll quickly learn how to deploy and then apply Sqoop in your environment. The authors provide MySQL, Oracle, and PostgreSQL database examples on GitHub that you can easily adapt for SQL Server, Netezza, Teradata, or other relational systems. Transfer data from a single database table into your Hadoop ecosystem Keep table data and Hadoop in sync by importing data incrementally Import data from more than one database table Customize transferred data by calling various database functions Export generated, processed, or backed-up data from Hadoop to your database Run Sqoop within Oozie, Hadoop's specialized workflow scheduler Load data into Hadoop's data warehouse (Hive) or database (HBase) Handle installation, connection, and syntax issues common to specific database vendors

Learning Hadoop 2 - Garry Turkington 2015-02-13

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.

Apache Spark in 24 Hours, Sams Teach Yourself - Jeffrey Aven 2016-08-31

Apache Spark is a fast, scalable, and flexible open source distributed processing engine for big data systems and is one of the most active open source big data projects to date. In

just 24 lessons of one hour or less, Sams Teach Yourself Apache Spark in 24 Hours helps you build practical Big Data solutions that leverage Spark's amazing speed, scalability, simplicity, and versatility. This book's straightforward, step-by-step approach shows you how to deploy, program, optimize, manage, integrate, and extend Spark–now, and for years to come. You'll discover how to create powerful solutions encompassing cloud computing, real-time stream processing, machine learning, and more. Every lesson builds on what you've already learned, giving you a rock-solid foundation for real-world success. Whether you are a data analyst, data engineer, data scientist, or data steward, learning Spark will help you to advance your career or embark on a new career in the booming area of Big Data. Learn how to • Discover what Apache Spark does and how it fits into the Big Data landscape • Deploy and run Spark locally or in the cloud • Interact with Spark from the shell • Make the most of the Spark Cluster Architecture • Develop Spark applications with Scala and functional Python • Program with the Spark API, including transformations and actions • Apply practical data engineering/analysis approaches designed for Spark • Use Resilient Distributed Datasets (RDDs) for caching, persistence, and output • Optimize Spark solution performance • Use Spark with SQL (via Spark SQL) and with NoSQL (via Cassandra) • Leverage cutting-edge functional programming techniques • Extend Spark with streaming, R, and Sparkling Water • Start building Spark-based machine learning and graph-processing applications • Explore advanced messaging technologies, including Kafka • Preview and prepare for Spark's next generation of innovations

Instructions walk you through common questions, issues, and tasks; Q-and-As, Quizzes, and Exercises build and test your knowledge; "Did You Know?" tips offer insider advice and shortcuts; and "Watch Out!" alerts help you avoid pitfalls. By the time you're finished, you'll be comfortable using Apache Spark to solve a wide spectrum of Big Data problems.

Big Data Using Hadoop and Hive - Nitin Kumar 2021-03-24

This book is the basic guide for developers, architects, engineers, and anyone who wants to start leveraging the open-source software Hadoop and Hive to build distributed, scalable concurrent big data applications. Hive will be used for reading, writing, and managing the large, data set files. The book is a concise guide on getting started with an overall understanding on Apache Hadoop and Hive and how they work together to speed up development with minimal effort. It will refer to simple concepts and examples, as they are likely to be the best teaching aids. It will explain the logic, code, and configurations needed to build a successful, distributed, concurrent application, as well as the reason behind those decisions. FEATURES: Shows how to leverage the open-source software Hadoop and Hive to build distributed, scalable, concurrent big data applications Includes material on Hive architecture with various storage types and the Hive query language Features a chapter on big data and how Hadoop can be used to solve the changes around it Explains the basic Hadoop setup, configuration, and optimization

Monitoring with Ganglia - Matt Massie 2012-11-09

Written by Ganglia designers and maintainers, this book shows you how to collect and visualize metrics

from clusters, grids, and cloud infrastructures at any scale. Want to track CPU utilization from 50,000 hosts every ten seconds? Ganglia is just the tool you need, once you know how its main components work together. This hands-on book helps experienced system administrators take advantage of Ganglia 3.x. Learn how to extend the base set of metrics you collect, fetch current values, see aggregate views of metrics, and observe time-series trends in your data. You'll also examine real-world case studies of Ganglia installs that feature challenging monitoring requirements. Determine whether Ganglia is a good fit for your environment Learn how Ganglia's gmond and gmetad daemons build a metric collection overlay Plan for scalability early in your Ganglia deployment, with valuable tips and advice Take data visualization to a new level with gweb, Ganglia's web frontend Write plugins to extend gmond's metric-collection capability Troubleshoot issues you may encounter with a Ganglia installation Integrate Ganglia with the sFlow and Nagios monitoring systems Contributors include: Robert Alexander, Jeff Buchbinder, Frederiko Costa, Alex Dean, Dave Josephsen, Peter Phaal, and Daniel Pocock. Case study writers include: John Allspaw, Ramon Bastiaans, Adam Compton, Andrew Dibble, and Jonah Horowitz.

Professional Hadoop Solutions - Boris Lublinsky 2013-09-12

The go-to guidebook for deploying Big Data solutions withHadoop Today's enterprise architects need to understand how the Hadoopframeworks and APIs fit together, and how they can be integrated todeliver real-world solutions. This book is a practical, detailedguide to building and implementing those solutions, with code-

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

Big Data Analytics with Hadoop 3 - Sridhar Alla 2018-05-31

Explore big data concepts, platforms, analytics, and their applications using the power of Hadoop 3 Key Features Learn Hadoop 3 to build effective big data analytics solutions on-premise and on cloud Integrate Hadoop with other big data tools such as R, Python, Apache Spark, and Apache Flink Exploit big data using Hadoop 3 with real-world examples Book Description Apache Hadoop is the most popular platform for big data processing, and

can be combined with a host of other big data tools to build powerful analytics solutions. Big Data Analytics with Hadoop 3 shows you how to do just that, by providing insights into the software as well as its benefits with the help of practical examples. Once you have taken a tour of Hadoop 3's latest features, you will get an overview of HDFS, MapReduce, and YARN, and how they enable faster, more efficient big data processing. You will then move on to learning how to integrate Hadoop with the open source tools, such as Python and R, to analyze and visualize data and perform statistical computing on big data. As you get acquainted with all this, you will explore how to use Hadoop 3 with Apache Spark and Apache Flink for real-time data analytics and stream processing. In addition to this, you will understand how to use Hadoop to build analytics solutions on the cloud and an end-to-end pipeline to perform big data analysis using practical use cases. By the end of this book, you will be well-versed with the analytical capabilities of the Hadoop ecosystem. You will be able to build powerful solutions to perform big data analytics and get insight effortlessly. What you will learn Explore the new features of Hadoop 3 along with HDFS, YARN, and MapReduce Get well-versed with the analytical capabilities of Hadoop ecosystem using practical examples Integrate Hadoop with R and Python for more efficient big data processing Learn to use Hadoop with Apache Spark and Apache Flink for real-time data analytics Set up a Hadoop cluster on AWS cloud Perform big data analytics on AWS using Elastic Map Reduce Who this book is for Big Data Analytics with Hadoop 3 is for you if you are looking to build high-performance analytics solutions for your enterprise

or business using Hadoop 3's powerful features, or you're new to big data analytics. A basic understanding of the Java programming language is required.

Architecting Modern Data Platforms - Jan Kunigk 2018-12-05

There's a lot of information about big data technologies, but splicing these technologies into an end-to-end enterprise data platform is a daunting task not widely covered. With this practical book, you'll learn how to build big data infrastructure both on-premises and in the cloud and successfully architect a modern data platform. Ideal for enterprise architects, IT managers, application architects, and data engineers, this book shows you how to overcome the many challenges that emerge during Hadoop projects. You'll explore the vast landscape of tools available in the Hadoop and big data realm in a thorough technical primer before diving into: Infrastructure: Look at all component layers in a modern data platform, from the server to the data center, to establish a solid foundation for data in your enterprise Platform: Understand aspects of deployment, operation, security, high availability, and disaster recovery, along with everything you need to know to integrate your platform with the rest of your enterprise IT Taking Hadoop to the cloud: Learn the important architectural aspects of running a big data platform in the cloud while maintaining enterprise security and high availability

Programming MapReduce with Scalding - Antonios Chalkiopoulos 2014-06-25

This book is an easy-to-understand, practical guide to designing, testing, and implementing complex MapReduce applications in Scala using the Scalding framework. It is packed with examples featuring log-processing, ad-targeting, and machine learning. This

book is for developers who are willing to discover how to effectively develop MapReduce applications. Prior knowledge of Hadoop or Scala is not required; however, investing some time on those topics would certainly be beneficial.

Big Data Forensics – Learning Hadoop Investigations - Joe Sremack 2015-09-24

Perform forensic investigations on Hadoop clusters with cutting-edge tools and techniques About This Book Identify, collect, and analyze Hadoop evidence forensically Learn about Hadoop's internals and Big Data file storage concepts A step-by-step guide to help you perform forensic analysis using freely available tools Who This Book Is For This book is meant for statisticians and forensic analysts with basic knowledge of digital forensics. They do not need to know Big Data Forensics. If you are an IT professional, law enforcement professional, legal professional, or a student interested in Big Data and forensics, this book is the perfect hands-on guide for learning how to conduct Hadoop forensic investigations. Each topic and step in the forensic process is described in accessible language. What You Will Learn Understand Hadoop internals and file storage Collect and analyze Hadoop forensic evidence Perform complex forensic analysis for fraud and other investigations Use state-of-the-art forensic tools Conduct interviews to identify Hadoop evidence Create compelling presentations of your forensic findings Understand how Big Data clusters operate Apply advanced forensic techniques in an investigation, including file carving, statistical analysis, and more In Detail Big Data forensics is an important type of digital investigation that involves the identification, collection, and

analysis of large-scale Big Data systems. Hadoop is one of the most popular Big Data solutions, and forensically investigating a Hadoop cluster requires specialized tools and techniques. With the explosion of Big Data, forensic investigators need to be prepared to analyze the petabytes of data stored in Hadoop clusters. Understanding Hadoop's operational structure and performing forensic analysis with court-accepted tools and best practices will help you conduct a successful investigation. Discover how to perform a complete forensic investigation of large-scale Hadoop clusters using the same tools and techniques employed by forensic experts. This book begins by taking you through the process of forensic investigation and the pitfalls to avoid. It will walk you through Hadoop's internals and architecture, and you will discover what types of information Hadoop stores and how to access that data. You will learn to identify Big Data evidence using techniques to survey a live system and interview witnesses. After setting up your own Hadoop system, you will collect evidence using techniques such as forensic imaging and application-based extractions. You will analyze Hadoop evidence using advanced tools and techniques to uncover events and statistical information. Finally, data visualization and evidence presentation techniques are covered to help you properly communicate your findings to any audience. Style and approach This book is a complete guide that follows every step of the forensic analysis process in detail. You will be guided through each key topic and step necessary to perform an investigation. Hands-on exercises are presented throughout the book, and technical reference guides and sample documents are

included for real-world use.

Hadoop Operations - Benjamin Poole 2014-11-03 Big data is a relative term describing a situation where the volume, velocity and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making . Big data is not a single technology but a combination of old and new technologies that helps companies gain actionable insight. Therefore, big data is the capability to manage a huge volume of disparate data, at the right speed, and within the right time frame to allow real-time analysis and reaction. As we note earlier in this chapter, big data is typically broken down by three characteristics: Volume: How much data Velocity: How fast that data is processed Variety: The various types of data Although it's convenient to simplify big data into the three Vs, it can be misleading and overly simplistic. For example, you may be managing a relatively small amount of very disparate, complex data or you may be processing a huge volume of very simple data. That simple data may be all structured or all unstructured. Even more important is the fourth V: veracity. How accurate is that data in predicting business value? Do the results of a big data analysis actually make sense? Determining relevant data is key to delivering value from massive amounts of data. However, big data is defined less by volume - which is a constantly moving target - than by its ever-increasing variety, velocity, variability and complexity .

Professional Hadoop - Benoy Antony 2016-05-03 The professional's one-stop guide to this open-source, Java-based big data framework Professional Hadoop is the complete reference and resource for experienced developers

looking to employ Apache Hadoop in real-world settings. Written by an expert team of certified Hadoop developers, committers, and Summit speakers, this book details every key aspect of Hadoop technology to enable optimal processing of large data sets. Designed expressly for the professional developer, this book skips over the basics of database development to get you acquainted with the framework's processes and capabilities right away. The discussion covers each key Hadoop component individually, culminating in a sample application that brings all of the pieces together to illustrate the cooperation and interplay that make Hadoop a major big data solution. Coverage includes everything from storage and security to computing and user experience, with expert guidance on integrating other software and more. Hadoop is quickly reaching significant market usage, and more and more developers are being called upon to develop big data solutions using the Hadoop framework. This book covers the process from beginning to end, providing a crash course for professionals needing to learn and apply Hadoop quickly. Configure storage, UE, and in-memory computing Integrate Hadoop with other programs including Kafka and Storm Master the fundamentals of Apache Big Top and Ignite Build robust data security with expert tips and advice Hadoop's popularity is largely due to its accessibility. Open-source and written in Java, the framework offers almost no barrier to entry for experienced database developers already familiar with the skills and requirements real-world programming entails. Professional Hadoop gives you the practical information and framework-specific skills you need quickly.

Hadoop: The Definitive Guide - Tom White 2015-03-25

Get ready to unlock the power of your data. With the fourth edition of this comprehensive guide, youâ??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. Youâ??ll learn about recent changes to Hadoop, and explore new case studies on Hadoopâ??s role in healthcare systems and genomics data processing. Learn fundamental components such as MapReduce, HDFS, and YARN Explore MapReduce in depth, including steps for developing applications with it Set up and maintain a Hadoop cluster running HDFS and MapReduce on YARN Learn two data formats: Avro for data serialization and Parquet for nested data Use data ingestion tools such as Flume (for streaming data) and Sqoop (for bulk data transfer) Understand how high-level data processing tools like Pig, Hive, Crunch, and Spark work with Hadoop Learn the HBase distributed database and the ZooKeeper distributed configuration service

Big Data and Hadoop - VK Jain 2017-01-01

This book introduces you to the Big Data processing techniques addressing but not limited to various BI (business intelligence) requirements, such as reporting, batch analytics, online analytical processing (OLAP), data mining and Warehousing, and predictive analytics. The book has been written on IBMs Platform of Hadoop

framework. IBM Infosphere BigInsight has the highest amount of tutorial matter available free of cost on Internet which makes it easy to acquire proficiency in this technique. This therefore becomes highly vunerable coaching materials in easy to learn steps. The book optimally provides the courseware as per MCA and M. Tech Level Syllabi of most of the Universities. All components of big Data Platform like Jaql, Hive Pig, Sqoop, Flume , Hadoop Streaming, Oozie: HBase, HDFS, FlumeNG, Whirr, Cloudera, Fuse , Zookeeper and Mahout: Machine learning for Hadoop has been discussed in sufficient Detail with hands on Exercises on each.

Monitoring Hadoop - Gurmukh Singh 2015-04-30

Mastering Hadoop - Sandeep Karanth 2014-12-29

Do you want to broaden your Hadoop skill set and take your knowledge to the next level? Do you wish to enhance your knowledge of Hadoop to solve challenging data processing problems? Are your Hadoop jobs, Pig scripts, or Hive queries not working as fast as you intend? Are you looking to understand the benefits of upgrading Hadoop? If the answer is yes to any of these, this book is for you. It assumes novice-level familiarity with Hadoop.

Virtualizing Hadoop - George Trujillo 2015-07-14

Plan and Implement Hadoop Virtualization for Maximum Performance, Scalability, and Business Agility

Enterprises running Hadoop must absorb rapid changes in big data ecosystems, frameworks, products, and workloads. Virtualized approaches can offer important advantages in speed, flexibility, and elasticity. Now, a world-class team of enterprise virtualization and big data experts guide you through the choices, considerations, and tradeoffs surrounding Hadoop virtualization. The

authors help you decide whether to virtualize Hadoop, deploy Hadoop in the cloud, or integrate conventional and virtualized approaches in a blended solution. First, Virtualizing Hadoop reviews big data and Hadoop from the standpoint of the virtualization specialist. The authors demystify MapReduce, YARN, and HDFS and guide you through each stage of Hadoop data management. Next, they turn the tables, introducing big data experts to modern virtualization concepts and best practices. Finally, they bring Hadoop and virtualization together, guiding you through the decisions you'll face in planning, deploying, provisioning, and managing virtualized Hadoop. From security to multitenancy to day-to-day management, you'll find reliable answers for choosing your best Hadoop strategy and executing it. Coverage includes the following: • Reviewing the frameworks, products, distributions, use cases, and roles associated with Hadoop • Understanding YARN resource management, HDFS storage, and I/O • Designing data ingestion, movement, and organization for modern enterprise data platforms • Defining SQL engine strategies to meet strict SLAs • Considering security, data isolation, and scheduling for multitenant environments • Deploying Hadoop as a service in the cloud • Reviewing the essential concepts, capabilities, and terminology of virtualization • Applying current best practices, guidelines, and key metrics for Hadoop virtualization • Managing multiple Hadoop frameworks and products as one unified system • Virtualizing master and worker nodes to maximize availability and performance • Installing and configuring Linux for a Hadoop environment

Expert Hadoop 2

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

applications in a Hadoop cluster Manage and protect Hadoop data and high availability Work with HDFS commands, file permissions, and storage management Move data, and use YARN to allocate resources and schedule jobs Manage job workflows with Oozie and Hue Secure, monitor, log, and optimize Hadoop Benchmark and troubleshoot Hadoop

Practical Hive - Scott Shaw 2016-08-27

Dive into the world of SQL on Hadoop and get the most out of your Hive data warehouses. This book is your go-to resource for using Hive: authors Scott Shaw, Ankur Gupta, David Kjerrumgaard, and Andreas Francois Vermeulen take you through learning HiveQL, the SQL-like language specific to Hive, to analyze, export, and massage the data stored across your Hadoop environment. From deploying Hive on your hardware or virtual machine and setting up its initial configuration to learning how Hive interacts with Hadoop, MapReduce, Tez and other big data technologies, Practical Hive gives you a detailed treatment of the software. In addition, this book discusses the value of open source software, Hive performance tuning, and how to leverage semi-structured and unstructured data. What You Will Learn Install and configure Hive for new and existing datasets Perform DDL operations Execute efficient DML operations Use tables, partitions, buckets, and user-defined functions Discover performance tuning tips and Hive best practices Who This Book Is For Developers, companies, and professionals who deal with large amounts of data and could use software that can efficiently manage large volumes of input. It is assumed that readers have the ability to work with SQL.

Hadoop Operations and Cluster Management

Cookbook - Shumin Guo 2013

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

APPLIED BIG DATA AND BUSINESS INTELLIGENCE WITH SOFTWARE TOOLS. - César Pérez López 2022-10-05

The book begins by looking at massive computing tools in Big Data ecosystems with a focus on Hadoop, Mapreduce, Hadoop Distribute File System, and Hadoop Common Components (Pig, Hive, Flume, Oozie, Hbase, Sqoop, Mahout, and others). Job automation and examples developed with SQL Server are discussed below. Apache Ambari's Hadoop ecosystem is also introduced. Additionally, the SAS Big Data Analytics tools are presented (SAS Access Interface to Hadoop, SAS Data Management, SAS Visual Analytics, SAS Visual Statistics, SAS In Memory Statistics for Hadoop, SAS High Performance Data Mining, SAS High Performance Text Mining, SAS VIYA, etc.) Big Data Analytics tools from Oracle (Big Data Appliance, Big Data Connectors, NoSQL Database, Exadata, Business Analytics, etc.), Microsoft (HDInsight, Azure, etc.) and IBM (IBM Solution for Hadoop Power Systems

Edition, IBM AIX Solution Editions for Cognos and SPSS, IBM SPSS Modeler, etc.). The quality and integrity of data in Big Data processes and the movement of data between clusters are addressed below. As an example, the copy and movement of databases between servers in SQL Server is developed. HYPER-V, Hadoop, and Ganglia cluster monitoring tools, as well as web interface and other tools, are covered later. Finally, the techniques of Big Data and Business Intelligence are deepened. The most important Business Intelligence tools (Business Objects, MicroStrategy, Tableau, Power BI, Qlik, Domo, Pentaho, etc.) are analyzed with special attention to dashboards. SAS Visual Analytics tools and SAP tools for dashboards are described. Finally, the implementation of KDD (Knowledge Discovery in Data Bases) with SAS (SAS Enterprise Miner) and IBM (IBM SPSS Modeler) tools is described through examples.

Beginning Apache Hadoop Administration - Prashant Nair 2017-09-07

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.

Topics in Performance Evaluation, Measurement and Characterization - Raghunath Nambiar 2012-08-04

This book constitutes the proceedings of the Third Technology Conference on Performance Evaluation and Benchmarking, TPCTC 2011, held in conjunction with the 37th International Conference on Very Large Data Bases, VLDB 2011, in Seattle, August/September 2011. The 12 full papers and 2 keynote papers were carefully selected and reviewed from numerous submissions. The papers present novel ideas and methodologies in performance evaluation, measurement, and characterization.

Data Analytics with Hadoop - Benjamin Bengfort 2016-06

Ready to use statistical and machine-learning techniques across large data sets? This practical guide shows you why the Hadoop ecosystem is perfect for the job. Instead of deployment, operations, or software development usually associated with distributed computing, you'll focus on particular analyses you can build, the data warehousing techniques that Hadoop provides, and higher order data workflows this framework can produce. Data scientists and analysts will learn how to perform a wide range of techniques, from writing MapReduce and Spark applications with Python to using advanced modeling and data management with Spark

MLlib, Hive, and HBase. You'll also learn about the analytical processes and data systems available to build and empower data products that can handle—and actually require—huge amounts of data. Understand core concepts behind Hadoop and cluster computing Use design patterns and parallel analytical algorithms to create distributed data analysis jobs Learn about data management, mining, and warehousing in a distributed context using Apache Hive and HBase Use Sqoop and Apache Flume to ingest data from relational databases Program complex Hadoop and Spark applications with Apache Pig and Spark DataFrames Perform machine learning techniques such as classification, clustering, and collaborative filtering with Spark's MLlib

Hadoop in 24 Hours, Sams Teach Yourself - Jeffrey Aven 2017-04-07

Apache Hadoop is the technology at the heart of the Big Data revolution, and Hadoop skills are in enormous demand. Now, in just 24 lessons of one hour or less, you can learn all the skills and techniques you'll need to deploy each key component of a Hadoop platform in your local environment or in the cloud, building a fully functional Hadoop cluster and using it with real programs and datasets. Each short, easy lesson builds on all that's come before, helping you master all of Hadoop's essentials, and extend it to meet your unique challenges. Apache Hadoop in 24 Hours, Sams Teach Yourself covers all this, and much more: Understanding Hadoop and the Hadoop Distributed File System (HDFS) Importing data into Hadoop, and process it there Mastering basic MapReduce Java programming, and using advanced MapReduce API concepts Making the most of Apache Pig and Apache Hive Implementing

and administering YARN Taking advantage of the full Hadoop ecosystem Managing Hadoop clusters with Apache Ambari Working with the Hadoop User Environment (HUE) Scaling, securing, and troubleshooting Hadoop environments Integrating Hadoop into the enterprise Deploying Hadoop in the cloud Getting started with Apache Spark Step-by-step instructions walk you through common questions, issues, and tasks; Q-and-As, Quizzes, and Exercises build and test your knowledge; "Did You Know?" tips offer insider advice and shortcuts; and "Watch Out!" alerts help you avoid pitfalls. By the time you're finished, you'll be comfortable using Apache Hadoop to solve a wide spectrum of Big Data problems.

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

Big Data Analytics with R and Hadoop is a tutorial style book that focuses on all the powerful big data tasks that can be achieved by integrating R and Hadoop.This book is ideal for R developers who are looking for a way to perform big data analytics with Hadoop. This book is also aimed at those who know Hadoop and want to build some intelligent applications over Big data with R packages. It would be helpful if readers have basic knowledge of R.

Hadoop For Dummies - Dirk deRoos 2014-04-14

Let Hadoop For Dummies help harness the power of your data and rein in the information overload Big data has become big business, and companies and organizations of all sizes are struggling to find ways to retrieve valuable information from their massive data sets with becoming overwhelmed. Enter Hadoop and this easy-to-understand For Dummies guide. Hadoop For Dummies helps readers understand the value of big data, make

a business case for using Hadoop, navigate the Hadoop ecosystem, and build and manage Hadoop applications and clusters. Explains the origins of Hadoop, its economic benefits, and its functionality and practical applications Helps you find your way around the Hadoop ecosystem, program MapReduce, utilize design patterns, and get your Hadoop cluster up and running quickly and easily Details how to use Hadoop applications for data mining, web analytics and personalization, large-scale text processing, data science, and problem-solving Shows you how to improve the value of your Hadoop cluster, maximize your investment in Hadoop, and avoid common pitfalls when building your Hadoop cluster From programmers challenged with building and maintaining affordable, scaleable data systems to administrators who must deal with huge volumes of information effectively and efficiently, this how-to has something to help you with Hadoop.

Mastering Hadoop 3 - Chanchal Singh 2019-02-28 A comprehensive guide to mastering the most advanced Hadoop 3 concepts Key FeaturesGet to grips with the newly introduced features and capabilities of Hadoop 3Crunch and process data using MapReduce, YARN, and a host of tools within the Hadoop ecosystemSharpen your Hadoop skills with real-world case studies and codeBook Description Apache Hadoop is one of the most popular big data solutions for distributed storage and for processing large chunks of data. With Hadoop 3, Apache promises to provide a high-performance, more fault-tolerant, and highly efficient big data processing platform, with a focus on improved scalability and increased efficiency. With this guide, you'll understand

advanced concepts of the Hadoop ecosystem tool. You'll learn how Hadoop works internally, study advanced concepts of different ecosystem tools, discover solutions to real-world use cases, and understand how to secure your cluster. It will then walk you through HDFS, YARN, MapReduce, and Hadoop 3 concepts. You'll be able to address common challenges like using Kafka efficiently, designing low latency, reliable message delivery Kafka systems, and handling high data volumes. As you advance, you'll discover how to address major challenges when building an enterprise-grade messaging system, and how to use different stream processing systems along with Kafka to fulfil your enterprise goals. By the end of this book, you'll have a complete understanding of how components in the Hadoop ecosystem are effectively integrated to implement a fast and reliable data pipeline, and you'll be equipped to tackle a range of real-world problems in data pipelines. What you will learnGain an in-depth understanding of distributed computing using Hadoop 3Develop enterprise-grade applications using Apache Spark, Flink, and moreBuild scalable and high-performance Hadoop data pipelines with security, monitoring, and data governanceExplore batch data processing patterns and how to model data in HadoopMaster best practices for enterprises using, or planning to use, Hadoop 3 as a data platformUnderstand security aspects of Hadoop, including authorization and authenticationWho this book is for If you want to become a big data professional by mastering the advanced concepts of Hadoop, this book is for you. You'll also find this book useful if you're a Hadoop professional looking to strengthen your knowledge

of the Hadoop ecosystem. Fundamental knowledge of the Java programming language and basics of Hadoop is necessary to get started with this book.

Big Data Processing With Hadoop - Revathi, T. 2018-11-16

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.