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

MADALYNN ANGELO

Hadoop 2.x Administration Cookbook "O'Reilly Media, Inc."
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

Hadoop 2 Quick-Start Guide Packt Publishing Ltd
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

Accumulo Packt Publishing Ltd

Learn how to quickly generate business intelligence, insights and create interactive dashboards for digital storytelling through various data sources with Redash Key Features Learn the best use of visualizations to build powerful interactive dashboards Create and share visualizations and data in your organization Work with different complexities of data from different data sources Book Description Data exploration and visualization is vital to Business Intelligence, the backbone of almost every enterprise or organization. Redash is a querying and visualization tool developed to simplify how marketing and business development departments are exposed to data. If you want to learn to create interactive dashboards with Redash, explore different visualizations, and share the insights with your peers, then this is the ideal book for you. The book starts with essential Business Intelligence concepts that are at the heart of data visualizations. You will learn how to find your way round Redash and its rich array of data visualization options for building interactive dashboards. You will learn how to create data storytelling and share these with peers. You will see how to connect to different data sources to process complex data, and then visualize this data to reveal valuable insights. By the end of this book, you will be confident with the Redash dashboarding tool to provide insight and communicate data storytelling. What you will learn Install Redash and troubleshoot installation errors Manage user roles and permissions Fetch data from various data sources Visualize and present data with Redash Create active alerts based on your data Understand Redash administration and customization Export, share and recount stories with Redash visualizations Interact programmatically with Redash through the Redash API Who this book is for This book is intended for Data Analysts, BI professionals and Data Developers, but can be useful to anyone who has a basic knowledge of SQL and a creative mind. Familiarity with basic BI concepts will be helpful, but no knowledge of Redash is required.

Apache Ignite Quick Start Guide Simon and Schuster

Now in its second edition, this book focuses on practical algorithms for mining data from even the largest datasets.

Practical Data Science with Hadoop and Spark "O'Reilly Media, Inc."

If you’re like most R users, you have deep knowledge and love for statistics. But as your organization continues to collect huge amounts of data, adding tools such as Apache Spark makes a lot of sense. With this practical book, data scientists and professionals working with large-scale data applications will learn how to use Spark from R to tackle big data and big compute problems. Authors Javier Luraschi, Kevin Kuo, and Edgar Ruiz show you how to use R with Spark to solve different data analysis problems. This book covers relevant data science topics, cluster computing, and issues that should interest even the most advanced users. Analyze, explore, transform, and visualize data in Apache Spark with R Create statistical models to extract information and predict outcomes; automate the process in production-ready workflows Perform analysis and modeling across many machines using distributed computing techniques Use large-scale data from multiple sources and different formats with ease from within Spark Learn about alternative modeling frameworks for graph processing, geospatial analysis, and genomics at scale Dive into advanced topics including custom transformations, real-time data processing, and creating custom Spark extensions

Cloudera Administration Handbook Pearson Education

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

The Complete Guide to Large-Scale Analysis and Modeling Addison-Wesley Professional

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

HBase Packt Publishing Ltd

Build efficient, high-performance & scalable systems to process large volumes of data with Apache Ignite Key Features Understand Apache Ignite's in-memory technology Create High-Performance app components with Ignite Build a real-time data streaming and complex event processing system Book Description Apache Ignite is a distributed in-memory platform designed to scale and process large volume of data. It can be integrated with microservices as well as monolithic systems, and can be used as a scalable, highly available and performant deployment platform for microservices. This book will teach you to use Apache Ignite for building a high-performance, scalable, highly available system architecture with data integrity. The book takes you through the basics of Apache Ignite and in-memory technologies. You will learn about installation and clustering Ignite nodes, caching topologies, and various caching strategies, such as cache aside, read and write through, and write behind. Next, you will delve into detailed aspects of Ignite’s data grid: web session clustering and querying data. You will learn how to process large volumes of data using compute grid and Ignite’s map-reduce and executor service. You will learn about the memory architecture of Apache Ignite and monitoring memory and caches. You will use Ignite for complex event processing, event streaming, and the time-series predictions of opportunities and threats. Additionally, you will go through off-heap and on-heap caching, swapping, and native and Spring framework integration with Apache Ignite. By the end of this book, you will be confident with all the features of Apache Ignite 2.x that can be used to build a high-performance system architecture. What you will learn Use Apache Ignite’s data grid and implement web session clustering Gain high performance and linear scalability with in-memory distributed data processing Create a microservice on top of Apache Ignite that can scale and perform Perform ACID-compliant CRUD operations on an Ignite cache Retrieve data from Apache Ignite’s data grid using SQL, Scan and Lucene Text query Explore complex event processing concepts and event streaming Integrate your Ignite app with the Spring framework Who this book is for The book is for Big Data professionals who want to learn the essentials of Apache Ignite. Prior experience in Java is necessary.

Mastering Hadoop 3 John Wiley & Sons

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.

Machine Learning with Scala Quick Start Guide Packt Publishing Ltd

Microsoft Azure HDInsight is Microsoft’s 100 percent compliant distribution of Apache Hadoop on Microsoft Azure. This means that standard Hadoop concepts and technologies apply, so learning the Hadoop stack helps you learn the HDInsight service. At the time of this writing, HDInsight (version 3.0) uses Hadoop version 2.2 and Hortonworks Data Platform 2.0. In Introducing Microsoft Azure HDInsight, we cover what big data really means, how you can use it to your advantage in your company or organization, and one of the services you can use to do that quickly–specifically, Microsoft’s HDInsight service. We start with an overview of big data and Hadoop, but we don’t emphasize only concepts in this book–we want you to jump in and get your hands dirty working with HDInsight in a practical way. To help you learn and even implement HDInsight right away, we focus on a specific use case that applies to almost any organization and demonstrate a process that you can follow along with. We also help you learn more. In the last chapter, we look ahead at the future of HDInsight and give you recommendations for self-learning so that you can dive deeper into important concepts and round out your education on working with big data.

Professional Hadoop Solutions Hadoop 2 Quick-Start GuideLearn the Essentials of Big Data Computing in the Apache Hadoop 2 Ecosystem

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

LiveLessons John Wiley & Sons

Supervised and unsupervised machine learning made easy in Scala with this quick-start guide. Key Features Construct and deploy machine learning systems that learn from your data and give accurate predictions Unleash the power of Spark ML along with popular machine learning algorithms to solve complex tasks in Scala. Solve hands-on problems by combining popular neural network architectures such as LSTM and CNN using Scala with DeepLearning4j library Book

Description Scala is a highly scalable integration of object-oriented nature and functional programming concepts that make it easy to build scalable and complex big data applications. This book is a handy guide for machine learning developers and data scientists who want to develop and train effective machine learning models in Scala. The book starts with an introduction to machine learning, while covering deep learning and machine learning basics. It then explains how to use Scala-based ML libraries to solve classification and regression problems using linear regression, generalized linear regression, logistic regression, support vector machine, and Naïve Bayes algorithms. It also covers tree-based ensemble techniques for solving both classification and regression problems. Moving ahead, it covers unsupervised learning techniques, such as dimensionality reduction, clustering, and recommender systems. Finally, it provides a brief overview of deep learning using a real-life example in Scala. What you will learn Get acquainted with JVM-based machine learning libraries for Scala such as Spark ML and Deeplearning4j Learn RDDs, DataFrame, and Spark SQL for analyzing structured and unstructured data Understand supervised and unsupervised learning techniques with best practices and pitfalls Learn classification and regression analysis with linear regression, logistic regression, Naïve Bayes, support vector machine, and tree-based ensemble techniques Learn effective ways of clustering analysis with dimensionality reduction techniques Learn recommender systems with collaborative filtering approach Delve into deep learning and neural network architectures Who this book is for This book is for machine learning developers looking to train machine learning models in Scala without spending too much time and effort. Some fundamental knowledge of Scala programming and some basics of statistics and linear algebra is all you need to get started with this book.

Hadoop Security Packt Publishing Ltd

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

Hadoop 2 Quick-start Guide "O'Reilly Media, Inc."

Web Scraping techniques are getting more popular, since data is as valuable as oil in 21st century. Through this book get some key knowledge about using XPath, regEX; web scraping libraries for R like rvest and RSelenium technologies. Key Features Techniques, tools and frameworks for web scraping with R Scrape data effortlessly from a variety of websites Learn how to selectively choose the data to scrape, and build your dataset Book Description Web scraping is a technique to extract data from websites. It simulates the behavior of a website user to turn the website itself into a web service to retrieve or introduce new data. This book gives you all you need to get started with scraping web pages using R programming. You will learn about the rules of RegEx and Xpath, key components for scraping website data. We will show you web scraping techniques, methodologies, and frameworks. With this book's guidance, you will become comfortable with the tools to write and test RegEx and XPath rules. We will focus on examples of dynamic websites for scraping data and how to implement the techniques learned. You will learn how to collect URLs and then create XPath rules for your first web scraping script using rvest library. From the data you collect, you will be able to calculate the statistics and create R plots to visualize them. Finally, you will discover how to use Selenium drivers with R for more sophisticated scraping. You will create AWS instances and use R to connect a PostgreSQL database hosted on AWS. By the end of the book, you will be sufficiently confident to create end-to-end web scraping systems using R. What you will learn Write and create regEX rules Write XPath rules to query your data Learn how web scraping methods work Use rvest to crawl web pages Store data retrieved from the web Learn the key uses of Rselenium to scrape data Who this book is for This book is for R programmers who want to get started quickly with web scraping, as well as data analysts who want to learn scraping using R. Basic knowledge of R is all you need to get started with this book.

Hadoop and Spark Fundamentals IGI Global

"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 for Developers and Administrators "O'Reilly Media, Inc."

Summary Hadoop in Practice, Second Edition provides over 100 tested, instantly useful techniques that will help you conquer big data, using Hadoop. This revised new edition covers changes and new features in the Hadoop core architecture, including MapReduce 2. Brand new chapters cover YARN and integrating Kafka, Impala, and Spark SQL with Hadoop. You'll also get new and updated techniques for Flume, Sqoop, and Mahout, all of which have seen major new versions recently. In short, this is the most practical, up-to-date coverage of Hadoop available anywhere. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Book It's always a good time to upgrade your Hadoop skills! Hadoop in Practice, Second Edition provides a collection of 104 tested, instantly useful techniques for analyzing real-time streams, moving data securely, machine learning, managing large-scale clusters, and taming big data using Hadoop. This completely revised edition covers changes and new features in Hadoop core, including MapReduce 2 and YARN. You'll pick up hands-on best practices for integrating Spark, Kafka, and Impala with Hadoop, and get new and updated techniques for the latest versions of Flume, Sqoop, and Mahout. In short, this is the most practical, up-to-date coverage of Hadoop available. Readers need to know a programming language like Java and have basic familiarity with Hadoop. What's Inside Thoroughly updated for Hadoop 2 How to write YARN applications Integrate real-time technologies like Storm, Impala, and Spark Predictive analytics using Mahout and RR Readers need to know a programming language like Java and have basic familiarity with Hadoop. About the Author Alex Holmes works on tough big-data problems. He is a software engineer, author, speaker, and blogger specializing in large-scale Hadoop projects. Table of Contents PART 1 BACKGROUND AND FUNDAMENTALS Hadoop in a heartbeat Introduction to YARN PART 2 DATA LOGISTICS Data serialization—working with text and beyond Organizing and optimizing data in HDFS Moving data into and out of Hadoop PART 3 BIG DATA PATTERNS Applying MapReduce patterns to big data Utilizing data structures and algorithms at scale Tuning, debugging, and testing PART 4 BEYOND MAPREDUCE SQL on Hadoop Writing a YARN application

Create and Share Interactive Dashboards Using Redash Packt Publishing Ltd

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

Apache Hadoop YARN Addison-Wesley Professional

Get up to speed on Apache Accumulo, the flexible, high-performance key/value store created by the National Security Agency (NSA) and based on Google's BigTable data storage system. Written by former NSA team members, this comprehensive tutorial and reference covers Accumulo architecture, application development, table design, and cell-level security. With clear information on system administration, performance tuning, and best practices, this book is ideal for developers seeking to write Accumulo applications, administrators charged with installing and maintaining Accumulo, and other professionals interested in what Accumulo has to offer. You will find everything you need to use this system fully. Get a high-level introduction to Accumulo's architecture and data model Take a rapid tour through single- and multiple-node installations, data ingest, and query Learn how to write Accumulo applications for several use cases, based on examples Dive into Accumulo internals, including information not available in the documentation Get detailed information for installing, administering, tuning, and measuring performance Learn best practices based on successful implementations in the field Find answers to common questions that every new Accumulo user asks

Apache Hadoop 3 Quick Start Guide Morgan Kaufmann

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.

Hadoop: The Definitive Guide Cambridge University Press

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