Where To Download Installing Hadoop 2 6 X On Windows 1 Free Download Pdf

Hadoop 2 Quick-Start Guide Hadoop 2.x Administration Cookbook ODROID Magazine Beginning Apache Hadoop Administration Hadoop Security Communication, Networks and Computing Mastering MongoDB 6.x Expert Hadoop 2 Administration Scaling Big Data with Hadoop and Solr - Second Edition Hadoop: Data Processing and Modelling Hadoop Real-World Solutions Cookbook Practical Hadoop Security Big Data Analytics with Hadoop 3 Hadoop in Action Practical Hadoop Ecosystem Big Data Analytics PySpark Recipes Pro Apache Hadoop Learning Hadoop 2 Apache Spark Graph Processing Knowledge Management in Organizations Fast Data Processing with Spark 2 Learn Hadoop in 24 Hours Apache Hadoop YARN Professional Hadoop Solutions Next Generation Databases Computational Life Sciences Spark in Action Mastering Machine Learning with Spark 2.x An Architecture for Fast and General Data Processing on Large Clusters Smart Trends in Information Technology and Computer Communications Big Data and The Internet of Things Scala Data Analysis Cookbook Data Algorithms Mastering MongoDB 3.x Hadoop: The Definitive Guide Hadoop Operations Apache Spark in 24 Hours, Sams Teach Yourself Professional Hadoop Practical Graph Analytics with Apache Giraph

Big Data and The Internet of Things Jun 26 2020 Enterprise Information Architecture for a New Age: Big Data and The Internet of Things, provides guidance in designing an information architecture to accommodate increasingly large amounts of data, massively large amounts of data, not only from traditional sources, but also from novel sources such everyday objects that are fast becoming wired into global Internet. No business can afford to be caught out by missing the value to be mined from the increasingly large amounts of available data generated by everyday devices. The text provides background as to how analytical solutions and enterprise architecture methodologies and concepts have evolved (including the roles of data warehouses, business intelligence tools, predictive analytics, data discovery, Big Data, and the impact of the Internet of Things). Then you’re taken through a series of steps by which to define a future state architecture and create a plan for how to reach that future state. Enterprise Information Architecture for a

New Age: Big Data and The Internet of Things helps you gain an understanding of the following: Implications of Big Data from a variety of new data sources (including data from sensors that are part of the Internet of Things) upon an information architecture How establishing a vision for data usage by defining a roadmap that aligns IT with line-of-business needs is a key early step The importance and details of taking a step-by-step approach when dealing with shifting business challenges and changing technology capabilities How to mitigate risk when evaluating existing infrastructure and designing and deploying new infrastructure Enterprise Information Architecture for a New Age: Big Data and The Internet of Things combines practical advice with technical considerations. Author Robert Stackowiak and his team are recognized worldwide for their expertise in large data solutions, including analytics. Don't miss your chance to read this book and gain the benefit of their advice as you look forward in thinking through your own choices and designing your own architecture to accommodate the burgeoning explosion in data that can be analyzed and converted into valuable information to drive your business forward toward success.

Next Generation Databases Jan 02 2021 "It's not easy to find such a generous book on big data and databases. Fortunately, this book is the one." Feng Yu. Computing Reviews. June 28, 2016. This is a book for enterprise architects, database administrators, and developers who need to understand the latest developments in database technologies. It is the book to help you choose the correct database technology at a time when concepts such as Big Data, NoSQL and NewSQL are making what used to be an easy choice into a complex decision with significant implications. The relational database (RDBMS) model completely dominated database technology for over 20 years. Today this "one size fits all" stability has been disrupted by a relatively recent explosion of new database technologies. These paradigm-busting technologies are powering the "Big Data" and "NoSQL" revolutions, as well as forcing fundamental changes in databases across the board. Deciding to use a relational database was once truly a no-brainer, and the various commercial relational databases competed on price, performance, reliability, and ease of use rather than on fundamental architectures. Today we are faced with choices between radically different database technologies. Choosing the right database today is a complex undertaking, with serious economic and technological consequences. Next Generation Databases demystifies today's new database technologies. The book describes what each technology was designed to solve. It shows how each technology can be used to solve real word application and business problems. Most importantly, this book highlights the architectural differences between technologies that are the critical factors to consider when choosing a database platform for new and upcoming projects. Introduces the new technologies that have revolutionized the database landscape Describes how each technology can be used to solve specific application or business challenges

Reviews the most popular new wave databases and how they use these new database technologies

ODROID Magazine Dec 25 2022 Table of Contents 6 Hacking Pokemon Go With an ODROID: How to Perform GPS Spoofing 10 Taking a Crack at Breaking WPA Networks - Part 2 14 ODROID-C1 Laptop: A Custom Home Project Codenamed “Redtop” 17 Pac-Man 256: A Classic Game? A New Twist on the Endless Runner Genre? Find Out! 18 Installing Hadoop and Spark Onto an ODROID-XU4 Cluster 22 Backup Scripts: Keep Your Data Safe For Your Peace of Mind 27 ODROID-C2 as an IoT Device: Interfacing With The Real World 31 Kodibuntu: Auto-Starting Kodi With a Full Ubuntu Distribution 32 A Car Computer For The Love of Customization: Chronicles of a Mad Scientist 35 Linux Gaming: Sega Saturn and CDEmu 39 The XU4 Punnet: A Printable Card Case for the ODROID-XU4 41 Why Does the Loser Seem to Touch the Finish Line First? Interesting Experiments to Understand the Difference of Shutter Mechanisms 42 ODROID-VU7 Plus: Your Favorite Touchscreen Now Offers Higher Resolution 43 Meet an ODROIDian: Radostan Riedel (@raybuntu), Talented LibreELEC Developer

Computational Life Sciences Dec 01 2020 This book broadly covers the given spectrum of disciplines in Computational Life Sciences, transforming it into a strong helping hand for teachers, students, practitioners and researchers. In Life Sciences, problem-solving and data analysis often depend on biological expertise combined with technical skills in order to generate, manage and efficiently analyse big data. These technical skills can easily be enhanced by good theoretical foundations, developed from well-chosen practical examples and inspiring new strategies. This is the innovative approach of Computational Life Sciences-Data Engineering and Data Mining for Life Sciences: We present basic concepts, advanced topics and emerging technologies, introduce algorithm design and programming principles, address data mining and knowledge discovery as well as applications arising from real projects. Chapters are largely independent and often flanked by illustrative examples and practical advise.

Professional Hadoop Solutions Feb 03 2021 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.

Mastering MongoDB 6.x Aug 21 2022 Design and build solutions with the most powerful document database, MongoDB Key Features Learn from the experts about every new feature in MongoDB 6 and 5 Develop applications and administer clusters using MongoDB on premise or in the cloud Explore code-rich case studies showcasing MongoDB's major features followed by best practices Book Description MongoDB is a leading non-relational database. This book covers all the major features of MongoDB including the latest version 6. MongoDB 6.x adds many new features and expands on existing ones such as aggregation, indexing, replication, sharding and MongoDB Atlas tools. Some of the MongoDB Atlas tools that you will master include Atlas dedicated clusters and Serverless, Atlas Search, Charts, Realm Application Services/Sync, Compass, Cloud Manager and Data Lake. By getting hands-on working with code using realistic use cases, you will master the art of modeling, shaping and querying your data and become the MongoDB oracle for the business. You will focus on broadly used and niche areas such as optimizing queries, configuring large-scale clusters, configuring your cluster for high performance and availability and many more. Later, you will become proficient in auditing, monitoring, and securing your clusters using a structured and organized approach. By the end of this book, you will have grasped all the practical understanding needed to design, develop, administer and scale MongoDB-based database applications both on premises and on the cloud. What you will learn Understand data modeling and schema design, including smart indexing Master querying data using aggregation Use distributed transactions, replication and sharding for better results Administer your database using backups and monitoring tools Secure your cluster with the best checklists and advice Master MongoDB Atlas, Search, Charts, Serverless, Realm, Compass, Cloud Manager and other tools offered in the cloud or on premises Integrate MongoDB with other big data sources Design and deploy MongoDB in mobile, IoT and serverless environments Who this book is for This book is for MongoDB developers and database administrators who want to learn how to model their data using MongoDB in depth, for both greenfield and existing projects. An understanding of MongoDB, shell command skills and basic database design concepts is required to get the most out of this book.

Communication, Networks and Computing Sep 22 2022

Data Algorithms Apr 24 2020 If you are ready to dive into the MapReduce framework for processing large datasets, this practical book takes you step by step through the algorithms and tools you need to build distributed MapReduce applications with Apache Hadoop or Apache Spark. Each chapter provides a recipe for solving a massive computational problem, such as building a recommendation system. You'll learn how to implement the appropriate MapReduce solution with code that you can use in your projects. Dr. Mahmoud Parsian covers basic design patterns, optimization techniques, and data mining and machine learning solutions for problems in bioinformatics, genomics, statistics, and social network analysis. This book also includes an overview of MapReduce, Hadoop, and Spark. Topics include: Market basket analysis for a large set of transactions Data mining algorithms (K-means, KNN, and Naive Bayes) Using huge genomic data to sequence DNA and RNA Naive Bayes theorem and Markov chains for data and market prediction Recommendation algorithms and pairwise document similarity Linear regression, Cox regression, and Pearson correlation Allelic frequency and mining DNA Social network analysis (recommendation systems, counting triangles, sentiment analysis)

Learning Hadoop 2 Aug 09 2021 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.

Hadoop Security Oct 23 2022 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.x Administration Cookbook Jan 26 2023 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

Apache Hadoop YARN Mar 04 2021 "Apache Hadoop is helping drive the Big Data revolution. Now, its data processing has been completely overhauled: Apache Hadoop YARN provides resource management at data center scale and easier ways to create distributed applications that process petabytes of data. And now in Apache HadoopTM YARN, two Hadoop technical leaders show you how to develop new applications and adapt existing code to fully leverage these revolutionary advances." -- From the Amazon

Smart Trends in Information Technology and Computer Communications Jul 28 2020 This book constitutes the refereed proceedings of the First International Conference on Smart Trends in Information Technology and Computer Communications, SmartCom 2016, held in Jaipur, India, in August 2016. The 106

revised papers presented were carefully reviewed and selected from 469 submissions. The papers address issues on smart and secure systems; technologies for digital world; data centric approaches; applications for e-agriculture and e-health; products and IT innovations; research for knowledge computing.
Beginning Apache Hadoop Administration Nov 24 2022 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.
Professional Hadoop Nov 19 2019 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.

Practical Hadoop Ecosystem Dec 13 2021 Learn how to use the Apache Hadoop projects, including MapReduce, HDFS, Apache Hive, Apache HBase, Apache Kafka, Apache Mahout, and Apache Solr. From setting up the environment to running sample applications each chapter in this book is a practical tutorial on using an Apache Hadoop ecosystem project. While several books on Apache Hadoop are available, most are based on the main projects, MapReduce and HDFS, and none discusses the other Apache Hadoop ecosystem projects and how they all work together as a cohesive big data development platform. What You Will Learn: Set up the environment in Linux for Hadoop projects using Cloudera Hadoop Distribution CDH 5 Run a MapReduce job Store data with Apache Hive, and Apache HBase Index data in HDFS with Apache Solr Develop a Kafka messaging system Stream Logs to HDFS with Apache Flume Transfer data from MySQL database to Hive, HDFS, and HBase with Sqoop Create a Hive table over Apache Solr Develop a Mahout User Recommender System Who This Book Is For: Apache Hadoop developers. Pre-requisite knowledge of Linux and some knowledge of Hadoop is required.

An Architecture for Fast and General Data Processing on Large Clusters Aug 29 2020 The past few years have seen a major change in computing systems, as growing data volumes and stalling processor speeds require more and more applications to scale out to clusters. Today, a myriad data sources, from the Internet to business operations to scientific instruments, produce large and valuable data streams. However, the processing capabilities of single machines have not kept up with the size of data. As a result, organizations increasingly need to scale out their computations over clusters. At the same time, the speed and sophistication required of data processing have grown. In addition to simple queries, complex algorithms like machine learning and graph analysis are becoming common. And in addition to batch processing, streaming analysis of real-time data is required to let organizations take timely action. Future computing platforms will need to not only scale out traditional workloads, but support these new applications too. This book, a revised version of the 2014 ACM Dissertation Award winning dissertation, proposes an architecture for cluster computing systems that can tackle emerging data processing workloads at scale. Whereas early cluster computing systems, like MapReduce, handled batch processing, our architecture also enables streaming and interactive queries, while keeping MapReduce's scalability and fault tolerance. And whereas most deployed systems only support simple one-pass computations (e.g., SQL queries), ours also extends to the multi-pass algorithms required for complex analytics like machine learning. Finally, unlike the specialized systems proposed for some of these workloads, our architecture allows these computations to be combined, enabling rich new applications that intermix, for example, streaming

and batch processing. We achieve these results through a simple extension to MapReduce that adds primitives for data sharing, called Resilient Distributed Datasets (RDDs). We show that this is enough to capture a wide range of workloads. We implement RDDs in the open source Spark system, which we evaluate using synthetic and real workloads. Spark matches or exceeds the performance of specialized systems in many domains, while offering stronger fault tolerance properties and allowing these workloads to be combined. Finally, we examine the generality of RDDs from both a theoretical modeling perspective and a systems perspective. This version of the dissertation makes corrections throughout the text and adds a new section on the evolution of Apache Spark in industry since 2014. In addition, editing, formatting, and links for the references have been added.

Apache Spark in 24 Hours, Sams Teach Yourself Dec 21 2019 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.

Learn Hadoop in 24 Hours Apr 05 2021 Hadoop has changed the way large data sets are analyzed, stored, transferred, and processed. At such low cost, it provides benefits like supports partial failure, fault tolerance, consistency, scalability, flexible schema, and so on. It also supports cloud computing. More and more number of individuals are looking forward to mastering their Hadoop skills. While initiating with Hadoop, most users are unsure about how to proceed with Hadoop. They are not aware of what are the pre-requisite or data structure they should be familiar with. Or How to make the most efficient use of Hadoop and its ecosystem. To help them with all these queries and other issues this e-book is designed. The book gives insights into many of Hadoop libraries and packages that are not known to many Big data Analysts and Architects. The e-book also tells you about Hadoop MapReduce and HDFS. The example in the e-book is well chosen and demonstrates how to control Hadoop ecosystem through various shell commands. With this book, users will gain expertise in Hadoop technology and its related components. The book leverages you with the best Hadoop content with the lowest price range. After going through this book, you will also acquire knowledge on Hadoop Security required for Hadoop Certifications like CCAH and CCDH. It is a definite guide to Hadoop. Table Of Content Chapter 1: What Is Big Data 1. Examples Of 'Big Data' 2. Categories Of 'Big Data' 3. Characteristics Of 'Big Data' 4. Advantages Of Big Data Processing Chapter 2: Introduction to Hadoop 1. Components of Hadoop 2. Features Of 'Hadoop' 3. Network Topology In Hadoop Chapter 3: Hadoop Installation Chapter 4: HDFS 1. Read Operation 2. Write Operation 3. Access HDFS using JAVA API 4. Access HDFS Using COMMAND-LINE INTERFACE Chapter 5: Mapreduce 1. How MapReduce works 2. How MapReduce Organizes Work? Chapter 6: First Program 1. Understanding MapReducer Code 2. Explanation of SalesMapper Class 3. Explanation of SalesCountryReducer Class 4. Explanation of SalesCountryDriver Class Chapter 7: Counters & Joins In MapReduce 1. Two types of counters 2. MapReduce Join Chapter 8: MapReduce Hadoop Program To Join Data Chapter 9: Flume and Sqoop 1. What is SQOOP in Hadoop? 2. What is FLUME in Hadoop? 3. Some Important features of FLUME Chapter 10: Pig 1. Introduction to PIG 2. Create your First PIG Program 3. PART 1) Pig Installation 4. PART 2) Pig Demo Chapter 11: OOZIE 1. What is OOZIE? 2. How does OOZIE work? 3. Example Workflow Diagram 4. Oozie workflow application 5. Why use Oozie? 6. FEATURES OF OOZIE

Spark in Action Oct 31 2020 Summary Spark in Action teaches you the theory and skills you need to effectively handle batch and streaming data using Spark. Fully updated for Spark 2.0. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Big data systems distribute datasets across clusters of machines, making it a

challenge to efficiently query, stream, and interpret them. Spark can help. It is a processing system designed specifically for distributed data. It provides easy-to-use interfaces, along with the performance you need for production-quality analytics and machine learning. Spark 2 also adds improved programming APIs, better performance, and countless other upgrades. About the Book Spark in Action teaches you the theory and skills you need to effectively handle batch and streaming data using Spark. You'll get comfortable with the Spark CLI as you work through a few introductory examples. Then, you'll start programming Spark using its core APIs. Along the way, you'll work with structured data using Spark SQL, process near-real-time streaming data, apply machine learning algorithms, and munge graph data using Spark GraphX. For a zero-effort startup, you can download the preconfigured virtual machine ready for you to try the book's code. What's Inside Updated for Spark 2.0 Real-life case studies Spark DevOps with Docker Examples in Scala, and online in Java and Python About the Reader Written for experienced programmers with some background in big data or machine learning. About the Authors Petar Ze?evi? and Marko Bona?i are seasoned developers heavily involved in the Spark community. Table of Contents PART 1 - FIRST STEPS Introduction to Apache Spark Spark fundamentals Writing Spark applications The Spark API in depth PART 2 - MEET THE SPARK FAMILY Sparkling queries with Spark SQL Ingesting data with Spark Streaming Getting smart with MLlib ML: classification and clustering Connecting the dots with GraphX PART 3 - SPARK OPS Running Spark Running on a Spark standalone cluster Running on YARN and Mesos PART 4 - BRINGING IT TOGETHER Case study: real-time dashboard Deep learning on Spark with H2O

Big Data Analytics Nov 12 2021 A handy reference guide for data analysts and data scientists to help to obtain value from big data analytics using Spark on Hadoop clusters About This Book This book is based on the latest 2.0 version of Apache Spark and 2.7 version of Hadoop integrated with most commonly used tools. Learn all Spark stack components including latest topics such as DataFrames, DataSets, GraphFrames, Structured Streaming, DataFrame based ML Pipelines and SparkR. Integrations with frameworks such as HDFS, YARN and tools such as Jupyter, Zeppelin, NiFi, Mahout, HBase Spark Connector, GraphFrames, H2O and Hivemall. Who This Book Is For Though this book is primarily aimed at data analysts and data scientists, it will also help architects, programmers, and practitioners. Knowledge of either Spark or Hadoop would be beneficial. It is assumed that you have basic programming background in Scala, Python, SQL, or R programming with basic Linux experience. Working experience within big data environments is not mandatory. What You Will Learn Find out and implement the tools and techniques of big data analytics using Spark on Hadoop clusters with wide variety of tools used with Spark and Hadoop Understand all the Hadoop and Spark ecosystem components Get to know all the Spark components:

Spark Core, Spark SQL, DataFrames, DataSets, Conventional and Structured Streaming, MLLib, ML Pipelines and Graphx See batch and real-time data analytics using Spark Core, Spark SQL, and Conventional and Structured Streaming Get to grips with data science and machine learning using MLLib, ML Pipelines, H2O, Hivemall, Graphx, SparkR and Hivemall. In Detail Big Data Analytics book aims at providing the fundamentals of Apache Spark and Hadoop. All Spark components – Spark Core, Spark SQL, DataFrames, Data sets, Conventional Streaming, Structured Streaming, MLlib, Graphx and Hadoop core components – HDFS, MapReduce and Yarn are explored in greater depth with implementation examples on Spark + Hadoop clusters. It is moving away from MapReduce to Spark. So, advantages of Spark over MapReduce are explained at great depth to reap benefits of in-memory speeds. DataFrames API, Data Sources API and new Data set API are explained for building Big Data analytical applications. Real-time data analytics using Spark Streaming with Apache Kafka and HBase is covered to help building streaming applications. New Structured streaming concept is explained with an IOT (Internet of Things) use case. Machine learning techniques are covered using MLLib, ML Pipelines and SparkR and Graph Analytics are covered with GraphX and GraphFrames components of Spark. Readers will also get an opportunity to get started with web based notebooks such as Jupyter, Apache Zeppelin and data flow tool Apache NiFi to analyze and visualize data. Style and approach This step-by-step pragmatic guide will make life easy no matter what your level of experience. You will deep dive into Apache Spark on Hadoop clusters through ample exciting real-life examples. Practical tutorial explains data science in simple terms to help programmers and data analysts get started with Data Science

Scala Data Analysis Cookbook May 26 2020 Navigate the world of data analysis, visualization, and machine learning with over 100 hands-on Scala recipes About This Book Implement Scala in your data analysis using features from Spark, Breeze, and Zeppelin Scale up your data anlytics infrastructure with practical recipes for Scala machine learning Recipes for every stage of the data analysis process, from reading and collecting data to distributed analytics Who This Book Is For This book shows data scientists and analysts how to leverage their existing knowledge of Scala for quality and scalable data analysis. What You Will Learn Familiarize and set up the Breeze and Spark libraries and use data structures Import data from a host of possible sources and create dataframes from CSV Clean, validate and transform data using Scala to pre-process numerical and string data Integrate quintessential machine learning algorithms using Scala stack Bundle and scale up Spark jobs by deploying them into a variety of cluster managers Run streaming and graph analytics in Spark to visualize data, enabling exploratory analysis In Detail This book will introduce you to the most popular Scala tools, libraries, and frameworks through practical recipes around loading, manipulating,

and preparing your data. It will also help you explore and make sense of your data using stunning and insightfulvisualizations, and machine learning toolkits. Starting with introductory recipes on utilizing the Breeze and Spark libraries, get to grips withhow to import data from a host of possible sources and how to pre-process numerical, string, and date data. Next, you'll get an understanding of concepts that will help you visualize data using the Apache Zeppelin and Bokeh bindings in Scala, enabling exploratory data analysis. iscover how to program quintessential machine learning algorithms using Spark ML library. Work through steps to scale your machine learning models and deploy them into a standalone cluster, EC2, YARN, and Mesos. Finally dip into the powerful options presented by Spark Streaming, and machine learning for streaming data, as well as utilizing Spark GraphX. Style and approach This book contains a rich set of recipes that covers the full spectrum of interesting data analysis tasks and will help you revolutionize your data analysis skills using Scala and Spark.

Hadoop 2 Quick-Start Guide Feb 27 2023 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

Knowledge Management in Organizations Jun 07 2021 This book contains the refereed proceedings of the 12th International Conference on Knowledge Management in Organizations, KMO 2017, held in Beijing, China, in August 2017. The theme of the conference was "Emerging Technology and Knowledge Management in Organizations." The 45 contributions accepted for KMO 2017 were selected from 112 submissions and are organized in topical sections on: Knowledge Management Models and Behaviour Studies; Knowledge Sharing; Knowledge Transfer and Learning; Knowledge and Service Innovation; Knowledge and Organization; Information Systems Research; Value Chain and Supply Chain; Knowledge Re-presentation and Reasoning; Data Mining and Intelligent Science; Big Data Management; Internet of Things and Network.

Mastering MongoDB 3.x Mar 24 2020 An expert's guide to build fault tolerant MongoDB application About This Book Master the advanced modeling, querying, and administration techniques in MongoDB and become a MongoDB expert Covers the latest updates and Big Data features frequently used by professional MongoDB developers and administrators If your goal is to become a certified MongoDB professional, this book is your perfect companion Who This Book Is For Mastering MongoDB is a book for database developers, architects, and administrators who want to learn how to use MongoDB more effectively and productively. If you have experience in, and are interested in working with, NoSQL databases to build apps and websites, then this book is for you. What You Will Learn Get hands-on with advanced querying techniques such as indexing, expressions, arrays, and more. Configure, monitor, and maintain highly scalable MongoDB environment like an expert. Master replication and data sharding to optimize read/write performance. Design secure and robust applications based on MongoDB. Administer MongoDB-based applications on-premise or in the cloud Scale MongoDB to achieve your design goals Integrate MongoDB with big data sources to process huge amounts of data In Detail MongoDB has grown to become the de facto NoSQL database with millions of users—from small startups to Fortune 500 companies. Addressing the limitations of SQL schema-based databases, MongoDB pioneered a shift of focus for DevOps and offered sharding and replication maintainable by DevOps teams. The book is based on MongoDB 3.x and covers topics ranging from database querying using the shell, built in drivers, and popular ODM mappers to more advanced topics such as sharding, high availability, and integration with big data sources. You will get an overview of MongoDB and how to play to its strengths, with relevant use cases. After that, you will learn how to query MongoDB effectively and make use of indexes as much as possible. The next part deals with the administration of MongoDB installations on-premise or in the cloud. We deal with database internals in the next section,

explaining storage systems and how they can affect performance. The last section of this book deals with replication and MongoDB scaling, along with integration with heterogeneous data sources. By the end this book, you will be equipped with all the required industry skills and knowledge to become a certified MongoDB developer and administrator. Style and approach This book takes a practical, step-by-step approach to explain the concepts of MongoDB. Practical use-cases involving real-world examples are used throughout the book to clearly explain theoretical concepts.

Pro Apache Hadoop Sep 10 2021 Pro Apache Hadoop, Second Edition brings you up to speed on Hadoop – the framework of big data. Revised to cover Hadoop 2.0, the book covers the very latest developments such as YARN (aka MapReduce 2.0), new HDFS high-availability features, and increased scalability in the form of HDFS Federations. All the old content has been revised too, giving the latest on the ins and outs of MapReduce, cluster design, the Hadoop Distributed File System, and more. This book covers everything you need to build your first Hadoop cluster and begin analyzing and deriving value from your business and scientific data. Learn to solve big-data problems the MapReduce way, by breaking a big problem into chunks and creating small-scale solutions that can be flung across thousands upon thousands of nodes to analyze large data volumes in a short amount of wall-clock time. Learn how to let Hadoop take care of distributing and parallelizing your software—you just focus on the code; Hadoop takes care of the rest. Covers all that is new in Hadoop 2.0 Written by a professional involved in Hadoop since day one Takes you quickly to the seasoned pro level on the hottest cloud-computing framework

Hadoop Operations Jan 22 2020 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

Hadoop in Action Jan 14 2022 Hadoop in Action teaches readers how to use Hadoop and write MapReduce programs. The intended readers are programmers,

architects, and project managers who have to process large amounts of data offline. Hadoop in Action will lead the reader from obtaining a copy of Hadoop to setting it up in a cluster and writing data analytic programs. The book begins by making the basic idea of Hadoop and MapReduce easier to grasp by applying the default Hadoop installation to a few easy-to-follow tasks, such as analyzing changes in word frequency across a body of documents. The book continues through the basic concepts of MapReduce applications developed using Hadoop, including a close look at framework components, use of Hadoop for a variety of data analysis tasks, and numerous examples of Hadoop in action. Hadoop in Action will explain how to use Hadoop and present design patterns and practices of programming MapReduce. MapReduce is a complex idea both conceptually and in its implementation, and Hadoop users are challenged to learn all the knobs and levers for running Hadoop. This book takes you beyond the mechanics of running Hadoop, teaching you to write meaningful programs in a MapReduce framework. This book assumes the reader will have a basic familiarity with Java, as most code examples will be written in Java. Familiarity with basic statistical concepts (e.g. histogram, correlation) will help the reader appreciate the more advanced data processing examples. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book.

Practical Hadoop Security Mar 16 2022 Practical Hadoop Security is an excellent resource for administrators planning a production Hadoop deployment who want to secure their Hadoop clusters. A detailed guide to the security options and configuration within Hadoop itself, author Bhushan Lakhe takes you through a comprehensive study of how to implement defined security within a Hadoop cluster in a hands-on way. You will start with a detailed overview of all the security options available for Hadoop, including popular extensions like Kerberos and OpenSSH, and then delve into a hands-on implementation of user security (with illustrated code samples) with both in-the-box features and with security extensions implemented by leading vendors. No security system is complete without a monitoring and tracing facility, so Practical Hadoop Security next steps you through audit logging and monitoring technologies for Hadoop, as well as ready to use implementation and configuration examples--again with illustrated code samples. The book concludes with the most important aspect of Hadoop security – encryption. Both types of encryptions, for data in transit and data at rest, are discussed at length with leading open source projects that integrate directly with Hadoop at no licensing cost. Practical Hadoop Security: Explains importance of security, auditing and encryption within a Hadoop installation Describes how the leading players have incorporated these features within their Hadoop distributions and provided extensions Demonstrates how to set up and use these features to your benefit and make your Hadoop installation secure without

impacting performance or ease of use

Mastering Machine Learning with Spark 2.x Sep 29 2020 Unlock the complexities of machine learning algorithms in Spark to generate useful data insights through this data analysis tutorial About This Book Process and analyze big data in a distributed and scalable way Write sophisticated Spark pipelines that incorporate elaborate extraction Build and use regression models to predict flight delays Who This Book Is For Are you a developer with a background in machine learning and statistics who is feeling limited by the current slow and “small data” machine learning tools? Then this is the book for you! In this book, you will create scalable machine learning applications to power a modern data-driven business using Spark. We assume that you already know the machine learning concepts and algorithms and have Spark up and running (whether on a cluster or locally) and have a basic knowledge of the various libraries contained in Spark. What You Will Learn Use Spark streams to cluster tweets online Run the PageRank algorithm to compute user influence Perform complex manipulation of DataFrames using Spark Define Spark pipelines to compose individual data transformations Utilize generated models for off-line/on-line prediction Transfer the learning from an ensemble to a simpler Neural Network Understand basic graph properties and important graph operations Use GraphFrames, an extension of DataFrames to graphs, to study graphs using an elegant query language Use K-means algorithm to cluster movie reviews dataset In Detail The purpose of machine learning is to build systems that learn from data. Being able to understand trends and patterns in complex data is critical to success; it is one of the key strategies to unlock growth in the challenging contemporary marketplace today. With the meteoric rise of machine learning, developers are now keen on finding out how can they make their Spark applications smarter. This book gives you access to transform data into actionable knowledge. The book commences by defining machine learning primitives by the MLlib and H2O libraries. You will learn how to use Binary classification to detect the Higgs Boson particle in the huge amount of data produced by CERN particle collider and classify daily health activities using ensemble Methods for Multi-Class Classification. Next, you will solve a typical regression problem involving flight delay predictions and write sophisticated Spark pipelines. You will analyze Twitter data with help of the doc2vec algorithm and K-means clustering. Finally, you will build different pattern mining models using MLlib, perform complex manipulation of DataFrames using Spark and Spark SQL, and deploy your app in a Spark streaming environment. Style and approach This book takes a practical approach to help you get to grips with using Spark for analytics and to implement machine learning algorithms. We'll teach you about advanced applications of machine learning through illustrative examples. These examples will equip you to harness the potential of machine learning, through Spark, in a variety of enterprise-grade systems.

Apache Spark Graph Processing Jul 08 2021 Build, process and analyze large-scale graph data effectively with Spark About This Book Find solutions for every stage of data processing from loading and transforming graph data to Improve the scalability of your graphs with a variety of real-world applications with complete Scala code. A concise guide to processing large-scale networks with Apache Spark. Who This Book Is For This book is for data scientists and big data developers who want to learn the processing and analyzing graph datasets at scale. Basic programming experience with Scala is assumed. Basic knowledge of Spark is assumed. What You Will Learn Write, build and deploy Spark applications with the Scala Build Tool. Build and analyze large-scale network datasets Analyze and transform graphs using RDD and graph-specific operations Implement new custom graph operations tailored to specific needs. Develop iterative and efficient graph algorithms using message aggregation and Pregel abstraction Extract subgraphs and use it to discover common clusters Analyze graph data and solve various data science problems using real-world datasets. In Detail Apache Spark is the next standard of open-source cluster-computing engine for processing big data. Many practical computing problems concern large graphs, like the Web graph and various social networks. The scale of these graphs - in some cases billions of vertices, trillions of edges - poses challenges to their efficient processing. Apache Spark GraphX API combines the advantages of both data-parallel and graph-parallel systems by efficiently expressing graph computation within the Spark data-parallel framework. This book will teach the user to do graphical programming in Apache Spark, apart from an explanation of the entire process of graphical data analysis. You will journey through the creation of graphs, its uses, its exploration and analysis and finally will also cover the conversion of graph elements into graph structures. This book begins with an introduction of the Spark system, its libraries and the Scala Build Tool. Using a hands-on approach, this book will quickly teach you how to install and leverage Spark interactively on the command line and in a standalone Scala program. Then, it presents all the methods for building Spark graphs using illustrative network datasets. Next, it will walk you through the process of exploring, visualizing and analyzing different network characteristics. This book will also teach you how to transform raw datasets into a usable form. In addition, you will learn powerful operations that can be used to transform graph elements and graph structures. Furthermore, this book also teaches how to create custom graph operations that are tailored for specific needs with efficiency in mind. The later chapters of this book cover more advanced topics such as clustering graphs, implementing graph-parallel iterative algorithms and learning methods from graph data. Style and approach A step-by-step guide that will walk you through the key ideas and techniques for processing big graph data at scale, with practical examples that will ensure an overall understanding of the concepts of Spark.

Fast Data Processing with Spark 2 May 06 2021 Learn how to use Spark to process big data at speed and scale for sharper analytics. Put the principles into practice for faster, slicker big data projects. About This Book A quick way to get started with Spark – and reap the rewards From analytics to engineering your big data architecture, we've got it covered Bring your Scala and Java knowledge – and put it to work on new and exciting problems Who This Book Is For This book is for developers with little to no knowledge of Spark, but with a background in Scala/Java programming. It's recommended that you have experience in dealing and working with big data and a strong interest in data science. What You Will Learn Install and set up Spark in your cluster Prototype distributed applications with Spark's interactive shell Perform data wrangling using the new DataFrame APIs Get to know the different ways to interact with Spark's distributed representation of data (RDDs) Query Spark with a SQL-like query syntax See how Spark works with big data Implement machine learning systems with highly scalable algorithms Use R, the popular statistical language, to work with Spark Apply interesting graph algorithms and graph processing with GraphX In Detail When people want a way to process big data at speed, Spark is invariably the solution. With its ease of development (in comparison to the relative complexity of Hadoop), it's unsurprising that it's becoming popular with data analysts and engineers everywhere. Beginning with the fundamentals, we'll show you how to get set up with Spark with minimum fuss. You'll then get to grips with some simple APIs before investigating machine learning and graph processing – throughout we'll make sure you know exactly how to apply your knowledge. You will also learn how to use the Spark shell, how to load data before finding out how to build and run your own Spark applications. Discover how to manipulate your RDD and get stuck into a range of DataFrame APIs. As if that's not enough, you'll also learn some useful Machine Learning algorithms with the help of Spark MLlib and integrating Spark with R. We'll also make sure you're confident and prepared for graph processing, as you learn more about the GraphX API. Style and approach This book is a basic, step-by-step tutorial that will help you take advantage of all that Spark has to offer.

Hadoop: Data Processing and Modelling May 18 2022 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.

Practical Graph Analytics with Apache Giraph Oct 19 2019 Practical Graph Analytics with Apache Giraph helps you build data mining and machine learning applications using the Apache Foundation’s Giraph framework for graph processing. This is the same framework as used by Facebook, Google, and other

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

Hadoop Real-World Solutions Cookbook Apr 17 2022 Over 90 hands-on recipes to help you learn and master the intricacies of Apache Hadoop 2.X, YARN, Hive, Pig, Oozie, Flume, Sqoop, Apache Spark, and Mahout About This Book Implement outstanding Machine Learning use cases on your own analytics models and processes. Solutions to common problems when working with the Hadoop ecosystem. Step-by-step implementation of end-to-end big data use cases. Who This Book Is For Readers who have a basic knowledge of big data systems and want to advance their knowledge with hands-on recipes. What You Will Learn Installing and maintaining Hadoop 2.X cluster and its ecosystem. Write advanced Map Reduce programs and understand design patterns. Advanced Data Analysis using the Hive, Pig, and Map Reduce programs. Import and export data from various sources using Sqoop and Flume. Data storage in various file formats such as Text, Sequential, Parquet, ORC, and RC Files. Machine learning principles with libraries such as Mahout Batch and Stream data processing using Apache Spark In Detail Big data is the current requirement. Most organizations produce huge amount of data every day. With the arrival of Hadoop-like tools, it has become easier for everyone to solve big data problems with great efficiency and at minimal cost. Grasping Machine Learning techniques will help you greatly in building predictive models and using this data to make the right decisions for your organization. Hadoop Real World Solutions Cookbook gives readers insights into learning and mastering big data via recipes. The book not only clarifies most big data tools in the market but also provides best practices for using them. The book provides recipes that are based on the latest versions of Apache Hadoop 2.X,

YARN, Hive, Pig, Sqoop, Flume, Apache Spark, Mahout and many more such ecosystem tools. This real-world-solution cookbook is packed with handy recipes you can apply to your own everyday issues. Each chapter provides in-depth recipes that can be referenced easily. This book provides detailed practices on the latest technologies such as YARN and Apache Spark. Readers will be able to consider themselves as big data experts on completion of this book. This guide is an invaluable tutorial if you are planning to implement a big data warehouse for your business. Style and approach An easy-to-follow guide that walks you through world of big data. Each tool in the Hadoop ecosystem is explained in detail and the recipes are placed in such a manner that readers can implement them sequentially. Plenty of reference links are provided for advanced reading.

Scaling Big Data with Hadoop and Solr - Second Edition Jun 19 2022 This book is aimed at developers, designers, and architects who would like to build big data enterprise search solutions for their customers or organizations. No prior knowledge of Apache Hadoop and Apache Solr/Lucene technologies is required.

Hadoop: The Definitive Guide Feb 21 2020 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 Analytics with Hadoop 3 Feb 15 2022 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.

Expert Hadoop 2 Administration Jul 20 2022 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

PySpark Recipes Oct 11 2021 Quickly find solutions to common programming problems encountered while processing big data. Content is presented in the popular problem-solution format. Look up the programming problem that you want to solve. Read the solution. Apply the solution directly in your own code. Problem solved! PySpark Recipes covers Hadoop and its shortcomings. The architecture of Spark, PySpark, and RDD are presented. You will learn to apply RDD to solve day-to-day big data problems. Python and NumPy are included and make it easy for new learners of PySpark to understand and adopt the model. What You Will Learn Understand the advanced features of PySpark2 and SparkSQL Optimize your code Program SparkSQL with Python Use Spark Streaming and Spark MLlib with Python Perform graph analysis with GraphFrames Who This Book Is For Data analysts, Python programmers, big data enthusiasts

- The Little Of Skin Care Korean Beauty Secrets For Healthy Glowing Skin
- Ams Weather Studies Investigations Manual Answer Key
- Mercury Grand Marquis Service Manual
- Holt Mcdougal Us History Teachers Edition
- Telling And Duxburys Planning Law And Procedure
- Finney Demana Waits Kennedy Calculus Graphical Numerical Algebraic 3rd Edition
- James C Livingston Anatomy Of The Sacred 6th Edition Book
- Sisters In The Wilderness Lives Of Susanna Moosie And Catharine Parr Traill Charlotte Gray
- Courageous Conversations About Race A Field Guide For Achieving Equity In Schools Glenn E Singleton
- Gendered Society Reader Kimmel 3rd Edition
- 35 The Endocrine System Study Guide Answers
- Criminal Justice An Introduction An Introduction To Crime And The Criminal Justice System
- Texas Staar Coach Math Workbooks
- Finish Line Mathematics Grade 7 Answer Key
- The Ancient Mysteries Of Melchizedek
- Guided Activity 4 1 Industrial Revolution Answers
- Lilley Pharmacology And The Nursing Process 6th Edition Test Bank
- Sida Badge Test Questions And Answers
- Ib Biology Questions And Answers
- From Poor Law To Welfare State A History Of Social In America Walter I Trattner
- Santrock Essentials Of Lifespan Development Mcgraw Hill
- Soluzioni Libro Prove Nazionali Matematica Spiga
- Mcgraw Hill Ehr Chapter
- Mastering The Teks In World History Answer Key Chapter 5
- Operations Research An Introduction 9th Edition Taha
- Dodge Durango Engine Diagram
- E2000 Manual User Guide
- Rawlinsons Construction Cost Guide Free
- Chapter 14 Section 3 Big Business Labor Answer Key
- Teaching From The Balance Point
- Doc Sloan Ritual Kappa Alpha Psi
- Reinforcement Activity 2 Part A Accounting Answers
- I Wish You More