Learning Spark

Learn how to hack systems like black hat hackers and secure them like security experts. Key Features Understand how computer systems work and their vulnerabilities Expose weaknesses and hack into machines to test their security Learn how to secure systems from hackers Book Description This book starts with the basics of ethical hacking, how to practice hacking safely and legally, and how to install and interact with Kali Linux and the Linux terminal. You will explore network hacking, where you will see how to test the security of wired and wireless networks. You will also learn how to crack the password for any Wi-Fi network (whether it uses WEP, WPA, or WPA2) and spy on the connected devices. Moving on, you will discover how to gain access to remote computer systems using client-side and server-side attacks. You will also get the hang of post-exploitation techniques, including remote control and interacting with the systems that you compromised. Towards the end of the book, you will be able to pick up web application hacking techniques. You'll see how to discover, exploit, and prevent a number of website vulnerabilities, such as XSS and SQL injections. The cover's attacks covered are practical techniques that work against real systems and are purely for educational purposes. At the end of each section, you will learn how to detect, prevent, and secure systems from these attacks. What you will learn Understand ethical hacking and the different fields and types of hackers Set up a penetration testing lab to practice safe and legal hacking Explore Linux basics, commands, and how to interact with the terminal Access password-protected networks and spy on connected devices and client-side attacks to hack and control remote computers Control a hacked system remotely and use it to hack other systems Discover, exploit, and prevent a number of web application vulnerabilities such as XSS and SQL injection. Who is this book for? Learning Ethical Hacking from Scratch is for anyone interested in learning how to hack and test the security of systems like professional hackers and security experts.

MongDB 4 Quick Start Guide

If your organization is about to enter the world of big data, you won't need to decide whether a pachce Hadoop is the right platform to use, but also which of its many components are best suited to your task. This field guide makes the exercise manageable by breaking down the Hadoop ecosystem into digestible sections. You'll quickly understand how Hadoop's projects, subprojects, and related technologies work together. Each chapter introduces a different topic—such as core technologies or data transfer—and explains why certain components may or may not be useful for particular needs. When it comes to data, Hadoop is a whole new ballgame, but with this handy reference, you'll have a good grasp of the playing field. Topics include: Core technologies—Hadoop Distributed File System (HDFS), MapReduce, YARN, and Spark Database and data management—Cassandra, HBase, MongoDB, and Hive Security, access control, auditing—Sentry, Kerberos, and Knox Cloud computing and virtualization—Senetig, Docker, and Hlinn Hadoop 2 Quick-start Guide

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 A lapati 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 size, spec, expand, and secure production Hadoop clusters of any size." — Paul Dix, Series Editor In Expert Hadoop Administration, leading "author and renowned Hadoop expert Sam A lapati 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 environments, A lapati 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. A lapati demystifies complex Hadoop environments, helping you understand exactly what happens behind the scenes when you administer your cluster. If you've pioneered unexplored data, your interest in Hadoop, or you've found"
Learn how to quickly generate business intelligence, insights and create interactive dashboards for digital storytelling through various techniques that complement it. Edaline concisely introduces and explains every key Hadoop 2 concept, tool, and service, illustrating each with a simple "beginning-to-end" example and identifying tradeoffs, 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 Edaline will bring you up to speed quickly, whether you’re a user, admin, developer, 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 RELMS 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, Oozie, and Hive OSSs Exploring applications, pipelines, and managing jobs, and making workflows managing Hadoop efficiently with Apache Air–including recipes for HDFS moving data in and out of Hadoop, and PART 3 BEYOND MAPREDUCE SQL on Hadoop Writing a YARN application function The following book introduces the usage of Apache Hadoop 2.x and YARN and the modern Hadoop ecosystem. Building on his unsurpassed experience teaching Hadoop and Big Data, author Douglas Edaline covers all the basics you need to know to install and use Hadoop 2.x on personal computers or servers, and to navigate the powerful parallel processing and data movement capabilities. If you are a big data enthusiast and striving to use Hadoop to solve your problems, this book is for you. Aimed at Java programmers with some knowledge of Hadoop MapReduce, this is also a comprehensive reference for developers and system administrators who want to get up to speed using Hadoop.

**Practical Data Science with Hadoop and Spark**

Ready to use statistical and machine-learning techniques across large data sets? This practical guide shows you why the Hadoop ecosystem is perfect for the job. Instead of deployment, operations, or software development usually associated with distributed computing, you’ll focus on particular analytic tools you can build, the data science and analytics workflows that Hadoop provides, and higher order data workflows this framework can produce. Data scientists and analysts will learn how to perform a wide range of techniques, from writing MapReduce and Spark applications with Python to using advanced modeling and data management with Spark MLlib, Hive, and HBase. You’ll also learn about the analytics and data science systems available to build and empower your data products that can handle— and actually require—huge amounts of data. Underlying Hadoop is a framework that enables you to ingest data patterns and parallel analytical algorithms to create distributed data analysis jobs. Learn about data management, mining, and warehousing in a distributed context using an Apache Hive and HBase database service. Snoop and a Apache Flume to ingest data from relational databases. Program complex Hadoop and Spark applications with a Apache Pig and Spark Databerames machine learning techniques such as classification, clustering, and collaborative filtering with Spark’s MLlib.

**MapReduce Design Patterns**

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 HBase, 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 e-book in PDF, Kindle, and ePub formats from Manning Publications. A bout the book: It’s always a good time to upgrade Spark. Kafka, and Spark SQL with Hadoop, and yet new and updated techniques for the latest versions of Flume, Sqoop, and HBase. 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 integrates real-time data processing engines like Storm, Impala, and Spark Predictive analytics using Python and R. 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 Modern World LOGISTICS Data serialization—working with text and beyond Optimizing and organizing data in HDFS Moving data in and out of Hadoop 3 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 function
Big Data Analytics with Hadoop 3

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 decision for using Hadoop, navigate the Hadoop ecosystem, and build and manage Hadoop applications and clusters. It explains the origins of Hadoop, its economic benefits, and its functionality and practical applications. It helps you find your way around the Hadoop ecosystem, program Hadoop, 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, scalable 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.

Next.js Quick Start Guide

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

Cloudera Administration Handbook

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 make 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 integrated object-oriented 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 Naive Bayes algorithms. It also covers tree-based ensembling techniques for solving both classification and regression problems. 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 will you learn? Get acquainted with JVM-based machine learning libraries for Scala such as Spark ML and Deeplearning4j. Learn RDOS, DataFrame, and Spark SQL for analyzing structured and unstructured data. Learn supervised and unsupervised learning techniques with best practices and pitfalls. Learn classification and regression analysis with linear regression, logistic regression, Naive Bayes, support vector machine, and tree-based ensembles. Learn applications of clustering analysis with dimensionality reduction techniques. Learn recommender systems with collaborative filtering approach Derive into deep learning network architectures Who is this book for? This is book 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.

Cloudera A dministration Handbook

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 it can offer you an 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.

Learn Ethical Hacking from Scratch

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 Explore big data using Hadoop 3 with real-world examples Book Description This book is the most popular platform for big data processing, and it 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. What will you learn? You will explore the new features of Hadoop 3 along with the MapReduce and Apache Flink well-versed with the analytical capabilities of Hadoop ecosystem using practical example in 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. A pache Flink is for you who 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. A pache Flink
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
- **Description**

A starter guide that covers Apache Flume in detail. Apache Flume: Distributed Log Collection for Hadoop is intended for people who are responsible for moving datasets into Hadoop in a timely and reliable manner like software engineers, database administrators, and data warehouse administrators.

**Hadoop Application Architectures**

The book contains the latest trend in IT industry—Big Data and Hadoop. It explains how big is 'Big Data' and why everybody is trying to implement this into their IT project. It includes research work on various topics, theoretical and practical approach, each component of the architecture keeps current industrial trends. Big Data and Hadoop have taken together are a new skill as per the industry standards. Readers will get a compact book along with the industry experience and would be a reference to help readers.

**Trino: The Definitive Guide**

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.

**IBM Data Engine for Hadoop and Spark**

"This text should be required reading for everyone in contemporary business." -- Peter Woodhall, CEO, M oda u2116 "The one book that clearly describes and links Big Data concepts to business utility." -- Dr. Christopher Starr, President "Simply, this is the best Big Data book on the market!" -- Sam Rodman, Cascadia IT Group "One of the most important books that I've seen on Big Data fundamentals." -- Joshua M. Davis, PhD

**Big Data and Hadoop**

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. **Description**

A starter guide that covers Apache Flume in detail. Apache Flume: Distributed Log Collection for Hadoop is intended for people who are responsible for moving datasets into Hadoop in a timely and reliable manner like software engineers, database administrators, and data warehouse administrators. 

**Big Data and Hadoop**

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. **Description**

A starter guide that covers Apache Flume in detail. Apache Flume: Distributed Log Collection for Hadoop is intended for people who are responsible for moving datasets into Hadoop in a timely and reliable manner like software engineers, database administrators, and data warehouse administrators.

**Hadoop Application Architectures**

The book contains the latest trend in IT industry—Big Data and Hadoop. It explains how big is 'Big Data' and why everybody is trying to implement this into their IT project. It includes research work on various topics, theoretical and practical approach, each component of the architecture keeps current industrial trends. Big Data and Hadoop have taken together are a new skill as per the industry standards. Readers will get a compact book along with the industry experience and would be a reference to help readers.
IoT Fundamentals

Today, billions of devices are Internet-connected, IoT standards and protocols are stabilizing, and technical professionals must increasingly solve the problems presented by the IoT technologies. Now, five leading Cisco IoT experts present the first comprehensive, practical reference for making IoT work. IoT Fundamentals brings together knowledge previously available only in white papers, standards documents, and other hard-to-find sources—on—or nowhere at all. The authors begin with a high-level overview of IoT and introduce key concepts needed to successfully design IoT solutions. They then work through each key technical IoT protocol, and technical building block that comprise complete IoT solutions. Building on these existing solutions, they present several detailed use cases, including manufacturing, energy, utilities, smart-connected cities, transportation, mining, and public safety. Whether you're a novice or existing infrastructure, you'll gain deep insight into how IoT applications are used. And components of next-generation wireless networks built with Cisco IoT solutions such as IEEE 802.11 (Wi-Fi), IEEE 802.15.4 (Mesh), and LORAWAN bring together real-world tips, insights, and best practices for designing and implementing next-generation wireless networks. Presenters start-to-finish configuration examples for common deployment scenarios reflect the extensive first-hand experience of Cisco experts.

Data-intensive Text Processing with MapReduce

This IBM Redbooks publication provides topics to help the technical community take advantage of the scalability, performance of the IBM Power Systems platform to implement or integrate an IBM Data Engine for Hadoop and Spark solution for analytics solutions that can access large data sets to improve business outcomes. This book documents topics to demonstrate and take advantage of the analytics strengths of the IBM POWER8® platform, the IBM analytics software portfolio, and selected third-party tools to help solve customer's data analytics workload requirements. This book describes how to plan, prepare, install, integrate, manage, and show how to use the IBM Data Engine for Hadoop and Spark solution to run analytic workloads on IBM POWER8®. In addition, this book delivers documentation to complement available IBM analytics solutions to help your data analytic needs. This publication strengthens the position of IBM analytics and big data solutions with a well-defined and documented deployment model within an IBM POWER™ virtualized environment so that customers have a planned foundation for security, scaling, capacity, resilience, and optimization for analytics workloads. This book is targeted at technical professionals (analytics consultants, technical support staff, IT Architects, and IT Specialists) that are responsible for delivering analytics solutions and support on IBM Power Systems.

Hadoop 2 Quick-Start Guide

Perform fast interactive analytics against different data sources using the T-Rino high-performance distributed SQL query engine. With this practical guide, you'll learn how to conduct analytics on data where it lives, whether it's Hive, Cassandra, a relational database, or a proprietary data store. A analyst, software engineers, and production engineers will learn how to manage, use, and even develop with T-Rino. Initially supported data source Trino is now used by Netflix, Airbnb, LinkedIn, Twitter, Uber, and many other companies. Matt Fuller, Manfred Moser, and Martin Traverso show you how to implement your own Trino query that can be used by many different sources to analyze across your entire organization. Get started: Explore Trino's use cases and learn about tools that will help you connect to Trino and query data Go-deeper: Learn Trino's features, including how to connect to and query data sources with support for SQL, statements, operators, functions, and more Put Trino in production: Secure Trino, monitor workloads, tune queries, and connect more applications; learn how other organizations apply Trino

Introducing Microsoft Azure HDInsight

Create, build and deploy universal JavaScript applications using Next.js: Key Features Work with the entire toolchain for developing universal JavaScript applications with Next.js: A straightforward guide to implementing server-side rendering with Next.js. Use Next.js to build SEO-friendly and superfast websites. Book Description Next.js is a powerful addition to the ever-growing and dynamic JavaScript world. Built on top of React, Webpack, and Babel, it is a minimalistic framework for server-rendered JavaScript applications. This book will show you the best practices for using Next.js, enabling you to build SEO-friendly and superfast websites. This book will guide you from building a simple single page app to a scalable and reliable client-server infrastructure. You will explore code sharing between client and server, universal modules, and server-side rendering. The book will take you through the next.js concepts that everyone is talking about today using Next.js: creating and using Next.js pages, server-side rendering, translation, CSS isolation, and more. You will learn ways of implementing them in order to create your own universal JavaScript application. You will walk through the building and deployment stages of your applications with the CD/ CI/ AP automation, including error handling, file downloading, deploying to production, and authentication. What you will learn The book will help you learn Next.js and its modules together by understanding code splitting and bundling Create website pages and wire them together with website navigation. Extend your application with additional features as custom Babel plugins and presets, React Router, Apollo, and GraphiQL. Next.js provides with Next.js to fetch data and receive push notifications Design and implement core modules, such as logging and authentication, and then more complex solutions for access control and business rule management Write tests and use online CI tools such as Travis, GitLab, and more. Build a Docker-based container for your app and deploy it to online services such as Heroku and Now. After this book is for JavaScript developers who want to learn how to generate server-rendered applications.

Big Data Fundamentals

A no-easy-to-follow Apache Hadoop administrator's guide filled with practical screenshots and explanations for each step and configuration. This book is great for administrators interested in setting up and managing a large Hadoop cluster. If you are an administrator, or want to be an administrator, and you are ready to build and maintain a production-level cluster running CDH5, then this book is for you.

Hadoop: The Definitive Guide

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 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 how to use Apache Ignite for building a high-performance, scalable, highly available system architecture with data integrity. You will take 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 cacheaside, 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 volume 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 cache. 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-line and on-line caching, and then learn about Apache Ignite's data grid and implement web session clustering. Gain high performance and linear scalability in distributed in-memory data processing Create a microservices framework that can scale and perform Perform ACID compliant CRUD operations on an in-memory 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 This book is for Big Data professionals who want to learn the essentials of Apache Ignite. Prior experience in Java is necessary.

Learning Hadoop 2

Our world is being revolutionized by data-driven methods: access to large amounts of data has generated new insights and opened exciting new opportunities in commerce, science, and computing applications. Processing the enormous quantities of data necessary for these advances requires large clusters, making distributed computing paradigms more crucial than ever. MapReduce is a programming model for expressing distributed computations on massive datasets and an execution framework for large-scale data processing on clusters of commodity servers.
Redash v5 Quick Start Guide

There is an easier way to build Hadoop applications. With this hands-on book, you'll learn how to use Cascading, the open source abstraction framework for Hadoop that lets you easily create and manage powerful enterprise-grade data processing applications—without having to learn the intricacies of MapReduce. Working with sample apps based on Java and other JVM languages, you'll quickly learn Cascading's streamlined approach to data processing, data filtering, and workflow optimization. This book demonstrates how this framework can help your business extract meaningful information from large amounts of distributed data. Start working on Cascading example projects right away and analyze unstructured data in any format, from any source and in test applications with familiar constructs and reusable components. Work with the Scalding and Cascalog Domain-Specific Languages. Easily deploy applications to Hadoop, regardless of cluster location or data size. Build workflows that integrate several big data frameworks and processes. Explore common use cases for Cascading, including features and tools that support them. Examine a case study that uses a dataset from the Open Data Initiative.

HBase Administration Cookbook

Get expert guidance on architecting end-to-end data management solutions with Apache Hadoop. While many sources explain how to use various components in the Hadoop ecosystem, this practical book takes you through architectural considerations necessary to tie those components together into a complete tailored application, based on your particular use case. To reinforce those lessons, the book's second section provides detailed examples of architectures used in some of the most commonly found Hadoop applications. Whether you're designing a new Hadoop application, or planning to integrate Hadoop into your existing data infrastructure, Hadoop Application Architectures will skillfully guide you through the process. This book covers: Factors to consider when using Hadoop to store and model data. Best practices for moving data in and out of the system. Data processing frameworks, including MapReduce, Spark, and Hive. Common Hadoop processing patterns, such as removing duplicate records and using windowing analytics. Giraph, GraphX, and other tools for large graph processing on Hadoop. Using workflow orchestration and scheduling tools such as a Apache Oozie. Near real-time stream processing with Apache Storm, Apache Spark Streaming, and Apache Flink. Architecture examples for clickstream analysis, fraud detection, and data warehousing.

HBase: The Definitive Guide

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 and regular expression (RegEx) in web scraping libraries for R like rvest and R Selenium 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. 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'll 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 the intricacies of web scraping with R for more sophisticated scraping. You will create AWS instances and use R to connect to 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. Use 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 Selenium to scrape data. Who is this book 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.