

Hadoop The Definitive Guide Tom White

Principles of Big Data helps readers avoid the common mistakes that endanger all Big Data projects. By stressing simple, fundamental concepts, this book teaches readers how to organize large volumes of complex data, and how to achieve data permanence when the content of the data is constantly changing. General methods for data verification and validation, as specifically applied to Big Data resources, are stressed throughout the book. The book demonstrates how adept analysts can find relationships among data objects held in disparate Big Data resources, when the data objects are endowed with semantic support (i.e., organized in classes of uniquely identified data objects). Readers will learn how their data can be integrated with data from other resources, and how the data extracted from Big Data resources can be used for purposes beyond those imagined by the data creators. Learn general methods for specifying Big Data in a way that is understandable to humans and to computers Avoid the pitfalls in Big Data design and analysis Understand how to create and use Big Data safely and responsibly with a set of laws, regulations and ethical standards that apply to the acquisition, distribution and integration of Big Data resources

A New York Times Bestseller, One of The Wall Street Journal's "10 Books to Read Now," and One of Kirkus Reviews's Best Nonfiction Books of Year We all sense it—something big is going on. You feel it in your workplace. You feel it when you talk to your kids. You can't miss it when you read the newspapers or watch the news. Our lives are being transformed in so many realms all at once—and it is dizzying. In Thank You for Being Late, version 2.0, with a new afterword, Thomas L. Friedman exposes the tectonic movements that are reshaping the world today and explains how to get the most out of them and cushion their worst impacts. His thesis: to understand the twenty-first century, you need to understand that the planet's three largest forces—Moore's law (technology), the Market (globalization), and Mother Nature (climate change and biodiversity loss)—are accelerating all at once. These accelerations are transforming five key realms: the workplace, politics, geopolitics, ethics, and community. The year 2007 was the major inflection point: the release of the iPhone, together with advances in silicon chips, software, storage, sensors, and networking, created a new technology platform that is reshaping everything from how we hail a taxi to the fate of nations to our most intimate relationships. It is providing vast new opportunities for individuals and small groups to save the world—or to destroy it. With his trademark vitality, wit, and optimism, Friedman shows that we can overcome the multiple stresses of an age of accelerations—if we slow down, if we dare to be late and use the time to reimagine work, politics, and community. Thank You for Being Late is an essential guide to the present and the future.

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

In addition to being an important component of the floras of Africa, Arabia and Madagascar, aloes are widely cultivated by hobbyists in Europe, the Americas, Australasia, Africa, India and other parts of the world. Aloe vera is one of the most important cultivated medicinal plants. Consequently they are of considerable importance in the horticultural trade and for their commercial value. Aloes: The definitive guide is a comprehensive account of all known aloes. For the first time since the publication of G.W. Reynolds' two books over 40 years ago, all currently accepted Aloe taxa are brought together, with descriptions, illustrations, reference topologies, location of types, distribution, habitat, and brief notes on diagnostic features and relationships. Dichotomous keys lead to identification of all taxa arranged in 10 groups based on habit and size, from small to large. An extensive, illustrated chapter on the history of discovery, especially in tropical Africa, adds interest. The text includes a pictorial glossary, bibliography, and index with synonyms. A must-have reference for succulent plant enthusiasts, hobbyists, nurserymen, CITES agencies, research workers and libraries, and botanists.

Hadoop: The Definitive Guide helps you harness the power of your data. Ideal for processing large datasets, the Apache Hadoop framework is an open source implementation of the MapReduce algorithm on which Google built its empire. This comprehensive resource demonstrates how to use Hadoop to build reliable, scalable, distributed systems: programmers will find details for analyzing large datasets, and administrators will learn how to set up and run Hadoop clusters. Complete with case studies that illustrate how Hadoop solves specific problems, this book helps you: Use the Hadoop Distributed File System (HDFS) for storing large datasets, and run distributed computations over those datasets using MapReduce Become familiar with Hadoop's data and I/O building blocks for compression, data integrity, serialization, and persistence Discover common pitfalls and advanced features for writing real-world MapReduce programs Design, build, and administer a dedicated Hadoop cluster, or run Hadoop in the cloud Use Pig, a high-level query language for large-scale data processing Take advantage of HBase, Hadoop's database for structured and semi-structured data Learn ZooKeeper, a toolkit of coordination primitives for building distributed systems If you have lots of data -- whether it's gigabytes or petabytes -- Hadoop is the perfect solution. Hadoop: The Definitive Guide is the most thorough book available on the subject. "Now you have the opportunity to learn about Hadoop from a master-not only of the technology, but also of common sense and plain talk."-- Doug Cutting, Hadoop Founder, Yahoo!

Ready to give your design skills a real boost? This eye-opening book helps you explore the design structure behind most of today's hit video games. You'll learn principles and practices for crafting games that generate emotionally charged experiences—a combination of elegant game mechanics, compelling fiction, and pace that fully immerses players. In clear and approachable prose, design pro Tynan Sylvester also looks at the day-to-day process necessary to keep your project on track, including how to work with a team,

and how to avoid creative dead ends. Packed with examples, this book will change your perception of game design. Create game mechanics to trigger a range of emotions and provide a variety of play Explore several options for combining narrative with interactivity Build interactions that let multiplayer gamers get into each other's heads Motivate players through rewards that align with the rest of the game Establish a metaphor vocabulary to help players learn which design aspects are game mechanics Plan, test, and analyze your design through iteration rather than deciding everything up front Learn how your game's market positioning will affect your design

"Get ready to unlock the power of your data. With the fourth edition of this comprehensive guide, you I 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."--

Data in all domains is getting bigger. How can you work with it efficiently? Recently updated for Spark 1.3, this book introduces Apache Spark, the open source cluster computing system that makes data analytics fast to write and fast to run. With Spark, you can tackle big datasets quickly through simple APIs in Python, Java, and Scala. This edition includes new information on Spark SQL, Spark Streaming, setup, and Maven coordinates. Written by the developers of Spark, this book will have data scientists and engineers up and running in no time. You'll learn how to express parallel jobs with just a few lines of code, and cover applications from simple batch jobs to stream processing and machine learning. Quickly dive into Spark capabilities such as distributed datasets, in-memory caching, and the interactive shell Leverage Spark's powerful built-in libraries, including Spark SQL, Spark Streaming, and MLlib Use one programming paradigm instead of mixing and matching tools like Hive, Hadoop, Mahout, and Storm Learn how to deploy interactive, batch, and streaming applications Connect to data sources including HDFS, Hive, JSON, and S3 Master advanced topics like data partitioning and shared variables For many organizations, Hadoop is the first step for dealing with massive amounts of data. The next step? Processing and analyzing datasets with the Apache Pig scripting platform. With Pig, you can batch-process data without having to create a full-fledged application, making it easy to experiment with new datasets. Updated with use cases and programming examples, this second edition is the ideal learning tool for new and experienced users alike. You'll find comprehensive coverage on key features such as the Pig Latin scripting language and the Grunt shell. When you need to analyze terabytes of data, this book shows you how to do it efficiently with Pig. Delve into Pig's data model, including scalar and complex data types Write Pig Latin scripts to sort, group, join, project, and filter your data Use Grunt to work with the Hadoop Distributed File System (HDFS) Build complex data processing pipelines with Pig's macros and modularity features Embed Pig Latin in Python for iterative processing and other advanced tasks Use Pig with Apache Tez to build high-performance batch and interactive data processing applications Create your own load and store functions to handle data formats and storage mechanisms 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. The programming model provides an easy-to-understand abstraction for designing scalable algorithms, while the execution framework transparently handles many system-level details, ranging from scheduling to synchronization to fault tolerance. This book focuses on MapReduce algorithm design, with an emphasis on text processing algorithms common in natural language processing, information retrieval, and machine learning. We introduce the notion of MapReduce design patterns, which represent general reusable solutions to commonly occurring problems across a variety of problem domains. This book not only intends to help the reader "think in MapReduce", but also discusses limitations of the programming model as well. This volume is a printed version of a work that appears in the Synthesis Digital Library of Engineering and Computer Science. Synthesis Lectures provide concise, original presentations of important research and development topics, published quickly, in digital and print formats. For more information visit www.morganclaypool.com Perform fast interactive analytics against different data sources using the Trino 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. Analysts, software engineers, and production engineers will learn how to manage, use, and even develop with Trino. Initially developed by Facebook, open 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 a single Trino query can combine data from multiple sources to allow for analytics 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 internal workings, 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

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

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 Model and analyze unstructured data in any format, from any source Build and 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 What could you do with data if scalability wasn't a problem? With this hands-on guide, you'll learn how Apache Cassandra handles hundreds of terabytes of data while remaining highly available across multiple data centers -- capabilities that have attracted Facebook, Twitter, and other data-intensive companies. Cassandra: The Definitive Guide provides the technical details and practical examples you need to assess this database management system and put it to work in a production environment. Author Eben Hewitt demonstrates the advantages of Cassandra's nonrelational design, and pays special attention

to data modeling. If you're a developer, DBA, application architect, or manager looking to solve a database scaling issue or future-proof your application, this guide shows you how to harness Cassandra's speed and flexibility. Understand the tenets of Cassandra's column-oriented structure Learn how to write, update, and read Cassandra data Discover how to add or remove nodes from the cluster as your application requires Examine a working application that translates from a relational model to Cassandra's data model Use examples for writing clients in Java, Python, and C# Use the JMX interface to monitor a cluster's usage, memory patterns, and more Tune memory settings, data storage, and caching for better performance

Let Hadoop For Dummies help harness the power of your data and rein in the information overload Big data has become big business, and companies and organizations of all sizes are struggling to find ways to retrieve valuable information from their massive data sets with becoming overwhelmed. Enter Hadoop and this easy-to-understand For Dummies guide. Hadoop For Dummies helps readers understand the value of big data, make a business case for using Hadoop, navigate the Hadoop ecosystem, and build and manage Hadoop applications and clusters. Explains the origins of Hadoop, its economic benefits, and its functionality and practical applications Helps you find your way around the Hadoop ecosystem, program MapReduce, utilize design patterns, and get your Hadoop cluster up and running quickly and easily Details how to use Hadoop applications for data mining, web analytics and personalization, large-scale text processing, data science, and problem-solving Shows you how to improve the value of your Hadoop cluster, maximize your investment in Hadoop, and avoid common pitfalls when building your Hadoop cluster From programmers challenged with building and maintaining affordable, 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.

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

Every enterprise application creates data, whether it's log messages, metrics, user activity, outgoing messages, or something else. And how to move all of this data becomes nearly as important as the data itself. If you're an application architect, developer, or production engineer new to Apache Kafka, this practical guide shows you how to use this open source streaming platform to handle real-time data feeds. Engineers from Confluent and LinkedIn who are responsible for developing Kafka explain how to deploy production Kafka clusters, write reliable event-driven microservices, and build scalable stream-processing applications with this platform. Through detailed examples, you'll learn Kafka's design principles, reliability guarantees, key APIs, and architecture details, including the replication protocol, the controller, and the storage layer. Understand publish-subscribe messaging and how it fits in the big data ecosystem. Explore Kafka producers and consumers for writing and reading messages Understand Kafka patterns and use-case requirements to ensure reliable data delivery Get best practices for building data pipelines and applications with Kafka Manage Kafka in production, and learn to perform monitoring, tuning, and maintenance tasks Learn the most critical metrics among Kafka's operational measurements Explore how Kafka's stream delivery capabilities make it a perfect source for stream processing systems

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

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

If you are a Big Data enthusiast and wish to use Hadoop v2 to solve your problems, then this book is for you. This book is for Java programmers with little to moderate knowledge of Hadoop MapReduce. This is also a one-stop reference for developers and system admins who want to quickly get up to speed with using Hadoop v2. It would be helpful to have a basic knowledge of software development using Java and a basic working knowledge of Linux.

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.

Hadoop: The Definitive Guide"O'Reilly Media, Inc."

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

Counsels programmers and administrators for big and small organizations on how to work with large-scale application datasets using Apache Hadoop, discussing its capacity for storing and processing large amounts of data while demonstrating best practices for building reliable and scalable distributed systems.

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

Data is bigger, arrives faster, and comes in a variety of formats—and it all needs to be processed at scale for analytics or machine learning. But how can you process such varied workloads efficiently? Enter Apache Spark. Updated to include Spark 3.0, this second edition shows data engineers and data scientists why structure and unification in Spark matters. Specifically, this book explains how to perform simple and complex data analytics and employ machine learning algorithms. Through step-by-step walk-throughs, code snippets, and notebooks, you'll be able to: Learn Python, SQL, Scala, or Java high-level Structured APIs Understand Spark operations and SQL Engine Inspect, tune, and debug Spark operations with Spark configurations and Spark UI Connect to data sources: JSON, Parquet, CSV, Avro, ORC, Hive, S3, or Kafka Perform analytics on batch and streaming data using Structured Streaming Build reliable data pipelines with open source Delta Lake and Spark Develop machine learning pipelines with MLlib and productionize models using MLflow

How can you get your data from frontend servers to Hadoop in near real time? With this complete reference guide, you'll learn Flume's rich set of features for collecting, aggregating, and writing large amounts of streaming data to the Hadoop Distributed File System (HDFS), Apache HBase, SolrCloud, Elastic Search, and other systems. Using Flume shows operations engineers how to configure, deploy,

and monitor a Flume cluster, and teaches developers how to write Flume plugins and custom components for their specific use-cases. You'll learn about Flume's design and implementation, as well as various features that make it highly scalable, flexible, and reliable. Code examples and exercises are available on GitHub. Learn how Flume provides a steady rate of flow by acting as a buffer between data producers and consumers Dive into key Flume components, including sources that accept data and sinks that write and deliver it Write custom plugins to customize the way Flume receives, modifies, formats, and writes data Explore APIs for sending data to Flume agents from your own applications Plan and deploy Flume in a scalable and flexible way—and monitor your cluster once it's running

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

If you're looking for a scalable storage solution to accommodate a virtually endless amount of data, this book shows you how Apache HBase can fulfill your needs. As the open source implementation of Google's BigTable architecture, HBase scales to billions of rows and millions of columns, while ensuring that write and read performance remain constant. Many IT executives are asking pointed questions about HBase. This book provides meaningful answers, whether you're evaluating this non-relational database or planning to put it into practice right away. Discover how tight integration with Hadoop makes scalability with HBase easier Distribute large datasets across an inexpensive cluster of commodity servers Access HBase with native Java clients, or with gateway servers providing REST, Avro, or Thrift APIs Get details on HBase's architecture, including the storage format, write-ahead log, background processes, and more Integrate HBase with Hadoop's MapReduce framework for massively parallelized data processing jobs Learn how to tune clusters, design schemas, copy tables, import bulk data, decommission nodes, and many other tasks

Until now, design patterns for the MapReduce framework have been scattered among various research papers, blogs, and books. This handy guide brings together a unique collection of valuable MapReduce patterns that will save you time and effort regardless of the domain, language, or development framework you're using. Each pattern is explained in context, with pitfalls and caveats clearly identified to help you avoid common design mistakes when modeling your big data architecture. This book also provides a complete overview of MapReduce that explains its origins and implementations, and why design patterns are so important. All code examples are written for Hadoop. Summarization patterns: get a top-level view by summarizing and grouping data Filtering patterns: view data subsets such as records generated from one user Data organization patterns: reorganize data to work with other systems, or to make MapReduce analysis easier Join patterns: analyze different datasets together to discover interesting relationships Metapatterns: piece together several patterns to solve multi-stage problems, or to perform several analytics in the same job Input and output patterns: customize the way you use Hadoop to load or store data "A clear exposition of MapReduce programs for common data processing patterns—this book is indispensable for anyone using Hadoop." --Tom White, author of Hadoop: The Definitive Guide

Data-driven insights are a key competitive advantage for any industry today, but deriving insights from raw data can still take days or weeks. Most organizations can't scale data science teams fast enough to keep up with the growing amounts of data to transform. What's the answer? Self-service data. With this practical book, data engineers, data scientists, and team managers will learn how to build a self-service data science platform that helps anyone in your organization extract insights from data. Sandeep Uttamchandani provides a scorecard to track and address bottlenecks that slow down time to insight across data discovery, transformation, processing, and production. This book bridges the gap between data scientists bottlenecked by engineering realities and data engineers unclear about ways to make self-service work. Build a self-service portal to support data discovery, quality, lineage, and governance Select the best approach for each self-service capability using open source cloud technologies Tailor self-service for the people, processes, and technology maturity of your data platform Implement capabilities to democratize data and reduce time to insight Scale your self-service portal to support a large number of users within your organization

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.

Get up to speed on the latest Ethernet capabilities for building and maintaining networks for everything from homes and offices to data centers and server machine rooms. This thoroughly revised, comprehensive guide covers a wide range of Ethernet technologies, from basic operation to network management, based on the authors' many years of field experience. When should you upgrade to higher speed Ethernet? How do you use switches to build larger networks? How do you troubleshoot the system? This book provides the answers. If you're looking to build a scalable network with Ethernet to satisfy greater bandwidth and market requirements, this book is indeed the definitive guide. Examine the most widely used media systems, as well as advanced 40 and 100 gigabit Ethernet Learn about Ethernet's four basic elements and the IEEE standards Explore full-duplex Ethernet, Power over Ethernet, and Energy Efficient Ethernet Understand structured cabling systems and the components you need to build your Ethernet system Use Ethernet switches to expand and improve network design Delve into Ethernet performance, from specific channels to the entire network Get troubleshooting techniques for problems common to twisted-pair and fiber optic systems

"Taking dynamic host and application metrics at scale"--Cover.

Learn how to use, deploy, and maintain Apache Spark with this comprehensive guide, written by the creators of the open-source cluster-computing framework. With an emphasis on improvements and new

features in Spark 2.0, authors Bill Chambers and Matei Zaharia break down Spark topics into distinct sections, each with unique goals. You'll explore the basic operations and common functions of Spark's structured APIs, as well as Structured Streaming, a new high-level API for building end-to-end streaming applications. Developers and system administrators will learn the fundamentals of monitoring, tuning, and debugging Spark, and explore machine learning techniques and scenarios for employing MLlib, Spark's scalable machine-learning library. Get a gentle overview of big data and Spark Learn about DataFrames, SQL, and Datasets—Spark's core APIs—through worked examples Dive into Spark's low-level APIs, RDDs, and execution of SQL and DataFrames Understand how Spark runs on a cluster Debug, monitor, and tune Spark clusters and applications Learn the power of Structured Streaming, Spark's stream-processing engine Learn how you can apply MLlib to a variety of problems, including classification or recommendation

Describes the features and functions of Apache Hive, the data infrastructure for Hadoop.

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

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

As we all know by now, wireless networks offer many advantages over fixed (or wired) networks. Foremost on that list is mobility, since going wireless frees you from the tether of an Ethernet cable at a desk. But that's just the tip of the cable-free iceberg. Wireless networks are also more flexible, faster and easier for you to use, and more affordable to deploy and maintain. The de facto standard for wireless networking is the 802.11 protocol, which includes Wi-Fi (the wireless standard known as 802.11b) and its faster cousin, 802.11g. With easy-to-install 802.11 network hardware available everywhere you turn, the choice seems simple, and many people dive into wireless computing with less thought and planning than they'd give to a wired network. But it's wise to be familiar with both the capabilities and risks associated with the 802.11 protocols. And 802.11 Wireless Networks: The Definitive Guide, 2nd Edition is the perfect place to start. This updated edition covers everything you'll ever need to know about wireless technology. Designed with the system administrator or serious home user in mind, it's a no-nonsense guide for setting up 802.11 on Windows and Linux. Among the wide range of topics covered are discussions on: deployment considerations network monitoring and performance tuning wireless security issues how to use and select access points network monitoring essentials wireless card configuration security issues unique to wireless networks With wireless technology, the advantages to its users are indeed plentiful. Companies no longer have to deal with the hassle and expense of wiring buildings, and households with several computers can avoid fights over who's online. And now, with 802.11 Wireless Networks: The Definitive Guide, 2nd Edition, you can integrate wireless technology into your current infrastructure with the utmost confidence.

[Copyright: 59f0d052ba1936e13fff32d0eda280ea](https://www.pdfdrive.com/hadoop-the-definitive-guide-tom-white.html)