

Pro Apache Hadoop

The book describes the emergence of big data technologies and the role of Spark in the entire big data stack. It compares Spark and Hadoop and identifies the shortcomings of Hadoop that have been overcome by Spark. The book mainly focuses on the in-depth architecture of Spark and our understanding of Spark RDDs and how RDD complements big data's immutable nature, and solves it with lazy evaluation, cacheable and type inference. It also addresses advanced topics in Spark, starting with the basics of Scala and the core Spark framework, and exploring Spark data frames, machine learning using Mlib, graph analytics using GraphX and real-time processing with Apache Kafka, AWS Kinesis, and Azure Event Hub. It then goes on to investigate Spark using PySpark and R. Focusing on the current big data stack, the book examines the interaction with current big data tools, with Spark being the core processing layer for all types of data. The book is intended for data engineers and scientists working on massive datasets and big data technologies in the cloud. In addition to industry professionals, it is helpful for aspiring data processing professionals and students working in big data processing and cloud computing environments.

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

Cybersecurity and Privacy in Cyber-Physical Systems collects and reports on recent high-quality research that addresses different problems related to cybersecurity and privacy in cyber-physical systems (CPSs). It Presents high-quality contributions addressing related theoretical and practical aspects Improves the reader's awareness of cybersecurity and privacy in CPSs Analyzes and presents the state of the art of CPSs, cybersecurity, and related technologies and methodologies Highlights and discusses recent developments and emerging trends in cybersecurity and privacy in CPSs Proposes new models, practical solutions, and technological advances related to cybersecurity and privacy in CPSs Discusses new cybersecurity and privacy models, prototypes, and protocols for CPSs This comprehensive book promotes high-quality research by bringing together researchers and experts in CPS security and privacy from around the world to share their knowledge of the different aspects of CPS security.

Cybersecurity and Privacy in Cyber-Physical Systems is ideally suited for policymakers, industrial engineers, researchers, academics, and professionals seeking a thorough understanding of the principles of cybersecurity and privacy in CPSs. They will learn about promising solutions to these research problems and identify unresolved and challenging problems for their own research. Readers will also have an overview of CPS cybersecurity and privacy design.

Big Data Systems encompass massive challenges related to data diversity, storage mechanisms, and requirements of massive computational power. Further, capabilities of big data systems also vary with respect to type of problems. For instance, distributed memory systems are not recommended for iterative algorithms. Similarly, variations in big data systems also exist related to consistency and fault tolerance. The purpose of this book is to provide a detailed explanation of big data systems. The book covers various topics including Networking, Security, Privacy, Storage, Computation, Cloud Computing, NoSQL and NewSQL systems, High Performance Computing, and Deep Learning. An illustrative and practical approach has been adopted in which theoretical topics have been aided by well-explained programming and illustrative examples. Key Features: Introduces concepts and evolution of Big Data technology. Illustrates examples for thorough understanding. Contains programming examples for hands on development. Explains a variety of topics including NoSQL Systems, NewSQL systems, Security, Privacy, Networking, Cloud, High Performance Computing, and Deep Learning. Exemplifies widely used big data technologies such as Hadoop and Spark. Includes discussion on case studies and open issues. Provides end of chapter questions for enhanced learning.

Pro MongoDB Development is about MongoDB, a NoSQL database based on the BSON (binary JSON) document model. The book discusses all aspects of using MongoDB in web applications: Java, PHP, Ruby, JavaScript are the most commonly used programming/scripting languages and the book discusses accessing MongoDB database with these languages. The book also discusses using Java EE frameworks Kundera and Spring Data with MongoDB. As NoSQL databases are commonly used with the Hadoop ecosystem the book also discusses using MongoDB with Apache Hive. Migration from other NoSQL databases (Apache Cassandra and Couchbase) and from relational databases (Oracle Database) is also discussed. What You'll Learn: How to use a Java client and MongoDB shell How to use MongoDB with PHP, Ruby, and Node.js as well How to migrate Apache Cassandra tables to MongoDB documents; Couchbase to MongoDB; and transferring data between Oracle and MongoDB How to use Kundera, Spring Data, and Spring XD with MongoDB How to load MongoDB data into Oracle Database and integrating MongoDB with Oracle Database in Oracle Data Integrator Audience: The target audience of the book is NoSQL database developers. Target audience includes Java, PHP and Ruby developers. The book is suitable for an intermediate level course in NoSQL database. Bring the power of functional programming to your PHP applications. From performance optimizations to concurrency, improved testability to code brevity, functional programming has a host of benefits when compared to traditional imperative programming. Part one of Pro Functional PHP Programming takes you through the basics of functional programming, outlining the key concepts and how they translate into standard PHP functions and code. Part two takes this theory and shows you the strategies for implementing it to solve real problems in your new or existing PHP applications. Functional programming is popular in languages

such as Lisp, Scheme and Clojure, but PHP also contains all you need to write functional code. This book will show you how to take advantage of functional programming in your own projects, utilizing the PHP programming language that you already know.

What You'll Learn Discover functional programming in PHP Work with functional programming functions Design strategies for high-performance applications Manage business logic with functions Use functional programming in object-oriented and procedural applications Employ helper libraries in your application Process big data with functional PHP Who This Book Is For Programmers and web developers with experience of PHP who are looking to get more out of their PHP coding and be able to do more with PHP.

Learn advanced analytical techniques and leverage existing tool kits to make your analytic applications more powerful, precise, and efficient. This book provides the right combination of architecture, design, and implementation information to create analytical systems that go beyond the basics of classification, clustering, and recommendation. Pro Hadoop Data Analytics emphasizes best practices to ensure coherent, efficient development. A complete example system will be developed using standard third-party components that consist of the tool kits, libraries, visualization and reporting code, as well as support glue to provide a working and extensible end-to-end system. The book also highlights the importance of end-to-end, flexible, configurable, high-performance data pipeline systems with analytical components as well as appropriate visualization results. You'll discover the importance of mix-and-match or hybrid systems, using different analytical components in one application. This hybrid approach will be prominent in the examples.

What You'll Learn Build big data analytic systems with the Hadoop ecosystem Use libraries, tool kits, and algorithms to make development easier and more effective Apply metrics to measure performance and efficiency of components and systems Connect to standard relational databases, noSQL data sources, and more Follow case studies with example components to create your own systems Who This Book Is For Software engineers, architects, and data scientists with an interest in the design and implementation of big data analytical systems using Hadoop, the Hadoop ecosystem, and other associated technologies.

In this age of data economy, data analytics is recognized as a key differentiator for companies trying to gain a sustainable competitive advantage and outperform their peers. However, the complexity of establishing an analytical architecture due to a wide array of disparate technical capabilities offered by a plethora of vendors makes the deployment of an on-premise solution a daunting task. For this reason, the Salesforce Analytics Cloud has captured the imagination from of both IT and business communities. The Salesforce Analytics Cloud represents the rethinking of analytics for the business user. The Analytics Cloud is a cloud-based platform designed for the business user to have access to analytics "on the go," providing answers to questions instantly on any device. This mobile-ready capability of the Salesforce Analytics Cloud means users can immediately collaborate and share insights with team members right inside Salesforce. Pro Salesforce Analytics Cloud provides actionable guidance on creating analytical capabilities using Salesforce Analytics Cloud. The book offers:

- A practical guide to Salesforce Analytics Cloud, including Wave Platform, Builder, and Explorer.
- Detailed business analytics use cases in various industries (e.g., retail, healthcare, financial services, and telecommunications).
- Architecture and best practices for integration, security, hybrid cloud

coexistence, and governance. *Pro Salesforce Analytics Cloud* is the first comprehensive book on Salesforce Analytics and provides a holistic architecture view of different analytical capabilities and how they fit into the overall information architecture. It features real-world industry use cases and demonstrates how Salesforce's Analytics Cloud solves business challenges and brings real value to the organization.

"The expert's voice in big data"--Cover.

"This book is a critically needed resource for the newly released Apache Hadoop 2.0, highlighting YARN as the significant breakthrough that broadens Hadoop beyond the MapReduce paradigm." —From the Foreword by Raymie Stata, CEO of Altiscale

The Insider's Guide to Building Distributed, Big Data Applications with Apache Hadoop™ YARN 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. YARN project founder Arun Murthy and project lead Vinod Kumar Vavilapalli demonstrate how YARN increases scalability and cluster utilization, enables new programming models and services, and opens new options beyond Java and batch processing. They walk you through the entire YARN project lifecycle, from installation through deployment. You'll find many examples drawn from the authors' cutting-edge experience—first as Hadoop's earliest developers and implementers at Yahoo! and now as Hortonworks developers moving the platform forward and helping customers succeed with it. Coverage includes YARN's goals, design, architecture, and components—how it expands the Apache Hadoop ecosystem

- Exploring YARN on a single node
- Administering YARN clusters and Capacity Scheduler
- Running existing MapReduce applications
- Developing a large-scale clustered YARN application
- Discovering new open source frameworks that run under YARN

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 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

This book covers three major parts of Big Data: concepts, theories and applications. Written by world-renowned leaders in Big Data, this book explores the problems, possible solutions and directions for Big Data in research and practice. It also focuses on high level concepts such as definitions of Big Data from different angles; surveys in research and applications; and existing tools, mechanisms, and systems in practice. Each chapter is independent from the other chapters, allowing users to read any chapter directly. After examining the practical side of Big Data, this book presents theoretical perspectives. The theoretical research ranges from Big Data representation, modeling and topology to distribution and dimension reducing. Chapters also investigate the many disciplines that involve Big Data, such as statistics, data mining, machine learning, networking, algorithms, security and differential geometry. The last section of this book introduces Big Data applications from different communities, such as business, engineering and science. Big Data Concepts, Theories and Applications is designed as a reference for researchers and advanced level students in computer science, electrical engineering and mathematics. Practitioners who focus on information systems, big data, data mining, business analysis and other related fields will also find this material valuable.

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

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. Since its release, Spring Framework has transformed virtually every aspect of Java development including web applications, security, aspect-oriented programming, persistence, and messaging. Spring Batch, one of its newer additions, now brings the same familiar Spring idioms to batch processing. Spring Batch addresses the needs of any batch process, from the complex calculations performed in the biggest financial institutions to simple data migrations that occur with many software development projects. Pro Spring Batch is intended to answer three questions: What? What is batch processing? What does it entail? What makes it different from the other applications we are developing? What are the challenges inherent in the development of a batch process? Why? Why do batch processing? Why can't we just process things as we get them? Why do we do batch processing differently than the web applications that we currently work on? How? How to implement a robust, scalable, distributed batch processing system using open-source frameworks Pro Spring Batch gives concrete examples of how each piece of functionality is used and why it would be used in a real-world application. This includes providing tips that the "school of hard knocks" has taught author Michael Minella during his experience with Spring Batch. Pro Spring Batch includes examples of I/O options that are not

mentioned in the official user's guide, as well as performance tips on things like how to limit the impact of maintaining the state of your jobs. The author also walks you through, from end to end, the design and implementation of a batch process based upon a theoretical real-world example. This includes basic project setup, implementation, testing, tuning and scaling for large volumes. 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

Data Scientists at Work is a collection of interviews with sixteen of the world's most influential and innovative data scientists from across the spectrum of this hot new profession. "Data scientist is the sexiest job in the 21st century," according to the Harvard Business Review. By 2018, the United States will experience a shortage of 190,000 skilled data scientists, according to a McKinsey report. Through incisive in-depth interviews, this book mines the what, how, and why of the practice of data science from the stories, ideas, shop talk, and forecasts of its preeminent practitioners across diverse industries: social network (Yann LeCun, Facebook); professional network (Daniel Tunkelang, LinkedIn); venture capital (Roger Ehrenberg, IA Ventures); enterprise cloud computing and neuroscience (Eric Jonas, formerly Salesforce.com); newspaper and media (Chris Wiggins, The New York Times); streaming television (Caitlin Smallwood, Netflix); music forecast (Victor Hu, Next Big Sound); strategic intelligence (Amy Heineike, Quid); environmental big data (André Karpištšenko, Planet OS); geospatial marketing intelligence (Jonathan Lenaghan, PlacelQ); advertising (Claudia Perlich, Dstillery); fashion e-commerce (Anna Smith, Rent the Runway); specialty retail (Erin Shellman, Nordstrom); email marketing (John Foreman, MailChimp); predictive sales intelligence (Kira Radinsky, SalesPredict); and humanitarian nonprofit (Jake Porway, DataKind). The book features a stimulating foreword by Google's Director of Research, Peter Norvig. Each of these data scientists shares how he or she tailors the torrent-taming techniques of big data, data visualization, search, and statistics to specific jobs by dint of ingenuity, imagination, patience, and passion. Data Scientists at Work parts the curtain on the interviewees' earliest data projects, how they became data scientists, their discoveries and surprises in working with data, their thoughts on the past, present, and future of the profession, their experiences of team collaboration within their organizations, and the insights they have gained as they get their hands dirty refining mountains of raw data into objects of commercial, scientific, and educational value for their organizations and clients.

In this fast-paced book on the Docker open standards platform for developing, packaging and running portable distributed applications, Deepak Vorhadiscusses how to build, ship and run applications on any platform such as a PC, the cloud, data center or a virtual machine. He describes how to install and create Docker images. and the advantages off Docker containers.The remainder of the book is devoted to

Read Free Pro Apache Hadoop

discussing using Docker with important software solutions. He begins by discussing using Docker with a traditional RDBMS using Oracle and MySQL. Next he moves on to NoSQL with chapter on MongoDB Cassandra, and Couchbase. Then he addresses the use of Docker in the Hadoop ecosystem with complete chapters on utilizing not only Hadoop, but Hive, HBase, Sqoop, Kafka, Solr and Spark. What You Will Learn How to install a Docker image How to create a Docker container How to run an Application in a Docker Container Use Docker with Apache Hadoop Ecosystem Use Docker with NoSQL Databases Use Docker with RDBMS Who This Book Is For Apache Hadoop Developers. Database developers. NoSQL Developers.

You've heard the hype about Hadoop: it runs petabyte-scale data mining tasks insanely fast, it runs gigantic tasks on clouds for absurdly cheap, it's been heavily committed to by tech giants like IBM, Yahoo!, and the Apache Project, and it's completely open-source (thus free). But what exactly is it, and more importantly, how do you even get a Hadoop cluster up and running? From Apress, the name you've come to trust for hands-on technical knowledge, Pro Hadoop brings you up to speed on Hadoop. You learn the ins and outs of MapReduce; how to structure a cluster, design, and implement the Hadoop file system; and how to build your first cloud-computing tasks using Hadoop. 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. Best of all, you'll learn from a tech professional who's been in the Hadoop scene since day one. Written from the perspective of a principal engineer with down-in-the-trenches knowledge of what to do wrong with Hadoop, you learn how to avoid the common, expensive first errors that everyone makes with creating their own Hadoop system or inheriting someone else's. Skip the novice stage and the expensive, hard-to-fix mistakes...go straight to seasoned pro on the hottest cloud-computing framework with Pro Hadoop. Your productivity will blow your managers away.

Pro Microsoft HDInsight is a complete guide to deploying and using Apache Hadoop on the Microsoft Windows Azure Platforms. The information in this book enables you to process enormous volumes of structured as well as non-structured data easily using HDInsight, which is Microsoft's own distribution of Apache Hadoop. Furthermore, the blend of Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) offerings available through Windows Azure lets you take advantage of Hadoop's processing power without the worry of creating, configuring, maintaining, or managing your own cluster. With the data explosion that is soon to happen, the open source Apache Hadoop Framework is gaining traction, and it benefits from a huge ecosystem that has risen around the core functionalities of the Hadoop distributed file system (HDFS™) and Hadoop Map Reduce. Pro Microsoft HDInsight equips you with the knowledge, confidence, and technique to configure and manage this ecosystem on Windows Azure. The book is an excellent choice for anyone aspiring to be a data scientist or data engineer, putting you a step ahead in the data mining field. Guides you through installation and configuration of an HDInsight cluster on Windows Azure Provides clear examples of configuring and executing Map Reduce jobs Helps you consume data and diagnose errors from the Windows Azure HDInsight Service

"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

Until recently, Hadoop deployments existed on hardware owned and run by organizations. Now, of course, you can acquire the computing resources and network connectivity to run Hadoop clusters in the cloud. But there's a lot more to deploying Hadoop to the public cloud than

simply renting machines. This hands-on guide shows developers and systems administrators familiar with Hadoop how to install, use, and manage cloud-born clusters efficiently. You'll learn how to architect clusters that work with cloud-provider features—not just to avoid pitfalls, but also to take full advantage of these services. You'll also compare the Amazon, Google, and Microsoft clouds, and learn how to set up clusters in each of them. Learn how Hadoop clusters run in the cloud, the problems they can help you solve, and their potential drawbacks. Examine the common concepts of cloud providers, including compute capabilities, networking and security, and storage. Build a functional Hadoop cluster on cloud infrastructure, and learn what the major providers require. Explore use cases for high availability, relational data with Hive, and complex analytics with Spark. Get patterns and practices for running cloud clusters, from designing for price and security to dealing with maintenance.

Leverage Phoenix as an ANSI SQL engine built on top of the highly distributed and scalable NoSQL framework HBase. Learn the basics and best practices that are being adopted in Phoenix to enable a high write and read throughput in a big data space. This book includes real-world cases such as Internet of Things devices that send continuous streams to Phoenix, and the book explains how key features such as joins, indexes, transactions, and functions help you understand the simple, flexible, and powerful API that Phoenix provides. Examples are provided using real-time data and data-driven businesses that show you how to collect, analyze, and act in seconds. Pro Apache Phoenix covers the nuances of setting up a distributed HBase cluster with Phoenix libraries, running performance benchmarks, configuring parameters for production scenarios, and viewing the results. The book also shows how Phoenix plays well with other key frameworks in the Hadoop ecosystem such as Apache Spark, Pig, Flume, and Sqoop. You will learn how to: Handle a petabyte data store by applying familiar SQL techniques. Store, analyze, and manipulate data in a NoSQL Hadoop ecosystem with HBase. Apply best practices while working with a scalable data store on Hadoop and HBase. Integrate popular frameworks (Apache Spark, Pig, Flume) to simplify big data analysis. Demonstrate real-time use cases and big data modeling techniques. Readership: Data Engineers, Big Data Administrators, and Architects.

Pro Website Development and Operations gives you the experience you need to create and operate a large-scale production website. Large-scale websites have their own unique set of problems regarding their design—problems that can get worse when agile methodologies are adopted for rapid results. Managing large-scale websites, deploying applications, and ensuring they are performing well often requires a full scale team involving the development and operations sides of the company—two departments that don't always see eye to eye. When departments struggle with each other, it adds unnecessary complexity to the work, and that result shows in the customer experience. Pro Website Development and Operations shows you how to streamline the work of web development and operations - incorporating the latest insights and methodologies of DevOps - so that your large-scale website is up and running quickly, with little friction and extreme efficiency between divisions. This book provides critical knowledge for any developer engaged in delivering the business and software engineering goals required to create and operate a large-scale production website. It addresses how developers can collaborate effectively with business and engineering teams to ensure applications are smoothly transitioned from product inception to implementation, and are properly deployed and managed. Pro Website

Development and Operations provides unique insights into how systems, code, and process can all work together to make large-scale website development and operations ultra-efficient.

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.

If your organization is about to enter the world of big data, you not only need to decide whether Apache 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 short, 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 Serialization—Avro, JSON, and Parquet Management and monitoring—Puppet, Chef, Zookeeper, and Oozie Analytic helpers—Pig, Mahout, and MLLib Data transfer—Scoop, Flume, distcp, and Storm Security, access control, auditing—Sentry, Kerberos, and Knox Cloud computing and virtualization—Serengeti, Docker, and Whirr

This book includes the outcomes of the International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD-2018), held in Tangier, Morocco on July 12–14, 2018. Presenting the latest research in the field of computing sciences and information technology, it discusses new challenges and provides valuable insights into the field, the goal being to stimulate debate, and to promote closer interaction and interdisciplinary collaboration between researchers and practitioners. Though chiefly intended for researchers and practitioners in advanced information

technology management and networking, the book will also be of interest to those engaged in emerging fields such as data science and analytics, big data, internet of things, smart networked systems, artificial intelligence, expert systems and cloud computing.

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

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

The go-to guidebook for deploying Big Data solutions with Hadoop Today's enterprise architects need to understand how the Hadoop frameworks and APIs fit together, and how they can be integrated to deliver real-world solutions. This book is a practical, detailed guide to building and implementing those solutions, with code-level instruction in the popular Wrox

tradition. It covers storing data with HDFS and Hbase, processing data with MapReduce, and automating data processing with Oozie. Hadoop security, running Hadoop with Amazon Web Services, best practices, and automating Hadoop processes in real time are also covered in depth. With in-depth code examples in Java and XML and the latest on recent additions to the Hadoop ecosystem, this complete resource also covers the use of APIs, exposing their inner workings and allowing architects and developers to better leverage and customize them. The ultimate guide for developers, designers, and architects who need to build and deploy Hadoop applications Covers storing and processing data with various technologies, automating data processing, Hadoop security, and delivering real-time solutions Includes detailed, real-world examples and code-level guidelines Explains when, why, and how to use these tools effectively Written by a team of Hadoop experts in the programmer-to-programmer Wrox style Professional Hadoop Solutions is the reference enterprise architects and developers need to maximize the power of Hadoop.

Pro Apache Hadoop

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

Pro SQL Database for Windows Azure, 2nd Edition shows how to create enterprise-level database deployments without the usual investment in datacenter and other infrastructure. Take advantage instead of Microsoft's worldwide backbone for cloud computing that delivers all the power of SQL Server in the form of the cloud-based SQL Database for Windows Azure. You can create and deploy a database in mere

minutes that is accessible worldwide and takes advantage of SQL Database's high-availability features to protect your data while ensuring 99.9% uptime. SQL Azure is ideally suited for startups, who can benefit from instant access to a robust and secure web-accessible database platform for use in rapidly deploying new products to market. SQL Azure is also ideal for small and mid-sized businesses, giving them the same ability to deploy SQL Server as any large enterprise, but without the management overhead. Even large enterprises find SQL Azure useful in creating failover environments, development environments, extra capacity to handle surges in demand, and more. Pro SQL Database for Windows Azure covers the very latest in Microsoft's fast-moving, cloud platform, showing how to program and administer it in a variety of cloud computing scenarios. You'll learn to program SQL Azure from ASP.NET, from WinForms, and from SQL Reporting Services. You'll learn to manage the platform by planning for scalability, troubleshooting performance issues, and implementing strong security. You'll learn the unique aspects of SQL Azure such as sharding and federation support that combine to place SQL Azure a step above and ahead of the competition. Shows how to use SQL Azure from classic Windows applications, ASP.NET and Windows Communication Foundation Covers management, performance, scalability, and troubleshooting Addresses the all-important issue of securing your data Helps you properly design for high-performance in a cloud environment Helps you adopt the new Federations feature in SQL Azure This book will be a step-by-step tutorial, which practically teaches working with big data on SQL Server through sample examples in increasing complexity. Microsoft SQL Server 2012 with Hadoop is specifically targeted at readers who want to cross-pollinate their Hadoop skills with SQL Server 2012 business intelligence and data analytics. A basic understanding of traditional RDBMS technologies and query processing techniques is essential.

Big Data: Principles and Paradigms captures the state-of-the-art research on the architectural aspects, technologies, and applications of Big Data. The book identifies potential future directions and technologies that facilitate insight into numerous scientific, business, and consumer applications. To help realize Big Data's full potential, the book addresses numerous challenges, offering the conceptual and technological solutions for tackling them. These challenges include life-cycle data management, large-scale storage, flexible processing infrastructure, data modeling, scalable machine learning, data analysis algorithms, sampling techniques, and privacy and ethical issues. Covers computational platforms supporting Big Data applications Addresses key principles underlying Big Data computing Examines key developments supporting next generation Big Data platforms Explores the challenges in Big Data computing and ways to overcome them Contains expert contributors from both academia and industry

The book 'Data Intensive Computing Applications for Big Data' discusses the technical concepts of big data, data intensive computing through machine learning, soft computing and parallel computing paradigms. It brings together researchers to report their latest results or progress in the development of the above mentioned areas. Since there are few books on this specific subject, the editors aim to provide a common platform for researchers working in this area to exhibit their novel findings. The book is intended as a reference work for advanced undergraduates and graduate students, as well as multidisciplinary, interdisciplinary and transdisciplinary research workers and scientists on the subjects of big data and cloud/parallel and distributed computing, and explains didactically many of the core concepts of these approaches for practical applications. It is organized into 24 chapters providing a comprehensive overview of big data analysis using parallel computing and addresses the complete data science workflow in the cloud, as well as dealing with privacy issues and the challenges faced in a data-intensive cloud computing environment. The book explores both fundamental and high-level concepts, and will serve as a manual for those in the industry, while also helping beginners to understand the basic and advanced aspects of big data and cloud computing.

Read Free Pro Apache Hadoop

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

Leverage Phoenix as an ANSI SQL engine built on top of the highly distributed and scalable NoSQL framework HBase. Learn the basics and best practices that are being adopted in Phoenix to enable a high write and read throughput in a big data space. This book includes real-world cases such as Internet of Things devices that send continuous streams to Phoenix, and the book explains how key features such as joins, indexes, transactions, and functions help you understand the simple, flexible, and powerful API that Phoenix provides. Examples are provided using real-time data and data-driven businesses that show you how to collect, analyze, and act in seconds. Pro Apache Phoenix covers the nuances of setting up a distributed HBase cluster with Phoenix libraries, running performance benchmarks, configuring parameters for production scenarios, and viewing the results. The book also shows how Phoenix plays well with other key frameworks in the Hadoop ecosystem such as Apache Spark, Pig, Flume, and Sqoop. You will learn how to: Handle a petabyte data store by applying familiar SQL techniques Store, analyze, and manipulate data in a NoSQL Hadoop echo system with HBase Apply best practices while working with a scalable data store on Hadoop and HBase Integrate popular frameworks (Apache Spark, Pig, Flume) to simplify big data analysis Demonstrate real-time use cases and big data modeling techniques Who This Book Is For Data engineers, Big Data administrators, and architects.

[Copyright: 1b48aa43ed33950e970a793394f81286](https://www.amazon.com/Pro-Apache-Phoenix-Data-Driven-Applications/dp/148aa43ed33950e970a793394f81286)