Cloudera Big Data Certification Toms It Pro

HBase is a remarkable tool for indexing mass volumes of data, but getting started with this distributed database and its ecosystem can be daunting. With this hands-on guide, you’ll learn how to architect, design, and deploy your own HBase applications by examining real-world solutions. Along with HBase principles and cluster deployment guidelines, this book includes in-depth case studies that demonstrate how large companies solved specific use cases with HBase. Authors Jean-Marc Spaggiari and Kevin O’Dell also provide draft solutions and code examples to help you implement your own versions of those use cases, from master data management (MDM) and document storage to near real-time event processing. You’ll also learn troubleshooting techniques to help you avoid common deployment mistakes. Learn exactly what HBase does, what its ecosystem includes, and how to set up your environment Explore how real-world HBase instances were deployed and put into production Examine documented use cases for tracking healthcare claims, digital advertising, data management, and product quality Understand how HBase works with tools and techniques such as Spark, Kafka, MapReduce, and the Java API Learn how to identify the causes and understand the consequences of the most common HBase issues

Now that people are aware that data can make the difference in an election or a business model, data science as an occupation is gaining ground. But how can you get started working in a wide-ranging, interdisciplinary field that’s so clouded in hype? This insightful book, based on Columbia University’s Introduction to Data Science class, tells you what you need to know. In many of these chapter-long lectures, data scientists from companies such as Google, Microsoft, and eBay share new algorithms, methods, and models by presenting case studies and the code they use. If you’re familiar with linear algebra, probability, and statistics, and have programming experience, this book is an ideal introduction to data science. Topics include: Statistical inference, exploratory data analysis, and the data science process Algorithms Spam filters, Naive Bayes, and data wrangling Logistic regression Financial modeling Recommendation engines and causality Data visualization Social networks and data journalism Data engineering, MapReduce, Pregel, and Hadoop Doing Data Science is collaboration between course instructor Rachel Schutt, Senior VP of Data Science at News Corp, and data science consultant Cathy O’Neil, a senior data scientist at Johnson Research Labs, who attended and blogged about the course.

Discover how Apache Hadoop can unleash the power of your data. This comprehensive resource shows you how to build and maintain reliable, scalable, distributed systems with the Hadoop framework -- an open source implementation of MapReduce, the algorithm on which Google built its empire. Programmers will find details for analyzing datasets of any size, and administrators will learn how to set up and run Hadoop clusters. This revised edition covers recent changes to Hadoop, including new features such as Hive, Sqoop, and Avro. It also provides illuminating case studies that illustrate how Hadoop is used to solve specific problems. Looking to get the most out of your data? This is your book. Use the Hadoop Distributed File System (HDFS) for storing large datasets, then run distributed computations over those datasets with 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 Analyze datasets with Hive, Hadoop’s data warehousing system Take advantage of HBase, Hadoop’s database for structured and semi-structured data Learn ZooKeeper, a toolkit of coordination primitives for building distributed systems "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, Cloudera

Grasp how mobile, big data, and analytics are combining to change business processes Right Experience, Right Results: Improving Profits, Margin, and Engagement with Mobile and Big Data illustrates how businesses can use mobility, big data, and analytics to enhance or change business processes, improve margins through better insight, transform customer experiences, empower employees with real-time, actionable insight, and more. The book depicts how companies can create competitive differentiation using mobile, cloud computing big data, and analytics to improve commerce, customer service, and communications with employees and consumers. In the past, the technologies used to deliver personalized and contextual services were either unavailable, unaffordable, or reserved solely for the consumer market. Today, however, the next wave of computing—mobile, cloud computing. big data, and analytics—has provided the foundation for businesses to create adaptive, personalized applications and services. Delivered point-of-need, these smarter services allow enterprise products and services to meet the burgeoning demand for always-connected, accurate, and real-time information. Right Experience, Right Results: Improving Profits, Margin, and Engagement with Mobile and Big Data is your guide to the new way of doing things. The book includes: Real world examples that illustrate how companies across various industries are creating better business processes by integrating new technologies A three step action plan for getting started and overcoming obstacles An electronic checklist with numerous actions that help get you up and running with incorporating mobile, big data, and analytics A guide to drawing insight from mobile, social, and other sources to create richer experiences If you're a CEO, chief marketing officer, marketing director, or business manager, Right Experience, Right Results gives you everything you need to harness technology to breathe new life into your business.

Secure your CISSP certification! If you’re a security professional seeking your CISSP certification, this book is a perfect way to prepare for the exam. Covering in detail all eight domains, the expert advice inside gives you the key information you'll need to pass the exam. Plus, you'll get tips on setting up a 60-day study plan, tips for exam day, and access to an online test bank of questions. CISSP For Dummies is fully updated and reorganized to reflect upcoming changes (ISC)2 has made to the Common Body of

Knowledge. Complete with access to an online test bank this book is the secret weapon you need to pass the exam and gain certification. Get key information for all eight exam domains Find test-taking and exam-day tips and tricks Benefit from access to free online practice questions and flash cards Prepare for the CISSP certification in 2018 and beyond You’ve put in the time as a security professional—and now you can reach your long-term goal of CISSP certification.

Blockchain technology is powering our future. As the technology behind cryptocurrencies like bitcoin and Facebook's Libra, open software platforms like Ethereum, and disruptive companies like Ripple, it’s too important to ignore. In this revelatory book, Don Tapscott, the bestselling author of Wikinomics, and his son, blockchain expert Alex Tapscott, bring us a brilliantly researched, highly readable, and essential book about the technology driving the future of the economy. Blockchain is the ingeniously simple, revolutionary protocol that allows transactions to be simultaneously anonymous and secure by maintaining a tamperproof public ledger of value. Though it’s best known as the technology that drives bitcoin and other digital currencies, it also has the potential to go far beyond currency, to record virtually everything of value to humankind, from birth and death certificates to insurance claims, land titles, and even votes. Blockchain is also essential to understand if you’re an artist who wants to make a living off your art, a consumer who wants to know where that hamburger meat really came from, an immigrant who’s tired of paying big fees to send money home to your loved ones, or an entrepreneur looking for a new platform to build a business. And those examples are barely the tip of the iceberg. As with major paradigm shifts that preceded it, blockchain technology will create winners and losers. This book shines a light on where it can lead us in the next decade and beyond.

Increase profits and reduce costs by utilizing this collectionof models of the most commonly asked data mining questions In order to find new ways to improve customer sales and support,and as well as manage risk, business managers must be able to minecompany databases. This book provides a step-by-step guide tocreating and implementing models of the most commonly asked datamining questions. Readers will learn how to prepare data to mine,and develop accurate data mining questions. The author, who hasover ten years of data mining experience, also provides actualtested models of specific data mining questions for marketing,sales, customer service and retention, and risk management. ACD-ROM, sold separately, provides these models for reader use.

Big Data in a nutshell: It is the ability to retain, process, and understand data like never before. It can mean more data than what you are using today; but it can also mean different kinds of data, a venture into the unstructured world where most of today's data resides. In this book you will learn how cognitive computing systems, like IBM Watson, fit into the Big Data world. Learn about the concept of data-in-motion and InfoSphere Streams, the world's fastest and most flexible platform for streaming data. Capturing, storing, refining, transforming, governing, securing, and analyzing data are important topics also covered in this book.

Boost your Big Data IQ! Gain insight into how to govern and consume IBM’s unique in-motion and at-rest Big Data analytic capabilities Big Data represents a new era of computing—an inflection point of opportunity where data in any format may be explored and utilized for breakthrough insights—whether that data is in-place, in-motion, or at-rest. IBM is uniquely positioned to help clients navigate this transformation. This book reveals how IBM is infusing open source Big Data technologies with IBM innovation that manifest in a platform capable of "changing the game." The four defining characteristics of Big Data—volume, variety, velocity, and veracity—are discussed. You’ll understand how IBM is fully committed to Hadoop and integrating it into the enterprise. Hear about how organizations are taking inventories of their existing Big Data assets, with search capabilities that help organizations discover what they could already know, and extend their reach into new data territories for unprecedented model accuracy and discovery. In this book you will also learn not just about the technologies that make up the IBM Big Data platform, but when to leverage its purpose-built engines for analytics on data in-motion and data at-rest. And you’ll gain an understanding of how and when to govern Big Data, and how IBM’s industry-leading InfoSphere integration and governance portfolio helps you understand, govern, and effectively utilize Big Data. Industry use cases are also included in this practical guide.

Fast data ingestion, serving, and analytics in the Hadoop ecosystem have forced developers and architects to choose solutions using the least common denominator—either fast analytics at the cost of slow data ingestion or fast data ingestion at the cost of slow analytics. There is an answer to this problem. With the Apache Kudu column-oriented data store, you can easily perform fast analytics on fast data. This practical guide shows you how. Begun as an internal project at Cloudera, Kudu is an open source solution compatible with many data processing frameworks in the Hadoop environment. In this book, current and former solutions professionals from Cloudera provide use cases, examples, best practices, and sample code to help you get up to speed with Kudu. Explore Kudu’s high-level design, including how it spreads data across servers Fully administer a Kudu cluster, enable security, and add or remove nodes Learn Kudu’s client-side APIs, including how to integrate Apache Impala, Spark, and other frameworks for data manipulation Examine Kudu’s schema design, including basic concepts and primitives necessary to make your project successful Explore case studies for using Kudu for real-time IoT analytics, predictive modeling, and in combination with another storage engine

Learn what it takes to succeed in the the most in-demand tech job Harvard Business Review calls it the sexiest tech job of the 21st century. Data scientists are in demand, and this unique book shows you exactly what employers want and the skill set that separates the quality data scientist from other talented IT professionals. Data science involves extracting, creating, and processing data to turn it into business value. With over 15 years of big data, predictive modeling, and business analytics experience, author Vincent Granville is no stranger to data science. In this one-of-a-kind guide, he provides insight into the essential data science skills, such as statistics and visualization techniques, and covers everything from analytical recipes and data science tricks to common job interview questions, sample resumes, and source code. The applications are endless and varied: automatically detecting spam and plagiarism, optimizing bid prices in keyword advertising, identifying new molecules to fight cancer, assessing the risk of meteorite impact. Complete with case studies, this book is a must, whether you're looking to become a data scientist or to hire one. Explains the finer points of data science, the required skills, and how to acquire them, including analytical recipes, standard rules, source code, and a dictionary of terms Shows what companies are looking for and how the growing importance of big data has increased the demand for data scientists Features job interview questions, sample resumes, salary surveys, and examples of job ads Case studies explore how data science is used on Wall Street, in botnet detection, for online advertising, and in many other business-critical situations Developing Analytic Talent: Becoming a Data Scientist is essential reading for those aspiring to this hot career choice and for employers seeking the best candidates.

Hone your analytic talents and become part of the next big thing Getting a Big Data Job For Dummies is the ultimate guide to landing a position in one of the fastest-growing fields in the modern economy. Learn exactly what "big data" means, why it's so important across all industries, and how you can obtain one of the most sought-after skill sets of the decade. This book walks you through the process of identifying your ideal big data job, shaping the perfect resume, and nailing the interview, all in one easy-to-read guide. Companies from all industries, including finance, technology, medicine, and defense, are harnessing massive amounts of data to reap a competitive advantage. The demand for big data professionals is growing every year, and experts forecast an estimated 1.9 million additional U.S. jobs in big data by 2015. Whether your niche is developing the technology, handling the data, or analyzing the results, turning your attention to a career in big data can lead to a more secure, more lucrative career path. Getting a Big Data Job For Dummies provides an overview of the big data career arc, and then shows you how to get your foot in the door with topics like: The education you need to succeed The range of big data career path options An overview of major big data employers A plan to develop your job-landing strategy Your analytic inclinations may be your ticket to long-lasting success. In a highly competitive job market, developing your data skills can create a situation where you pick your employer rather than the other way around. If you're ready to get in on the ground floor of the next big thing, Getting a Big Data Job For Dummies will teach you everything you need to know to get started today.

Prepare for the SAS Base Programming for SAS 9 exam with the official guide by the SAS Global Certification Program. New and experienced SAS users who want to prepare for the SAS Base Programming for SAS 9 exam will find this guide to be an invaluable, convenient, and comprehensive resource that covers all of the objectives tested on the exam. Now in its fourth edition, the guide has been extensively updated, and revised to streamline explanations. Major topics include importing and exporting raw data files, creating and modifying SAS data sets, and identifying and correcting data syntax and programming logic errors. The chapter quizzes have been thoroughly updated and full solutions are included at the back of the book. In addition, links are provided to the exam objectives, practice exams, and other helpful resources, such as the updated Base SAS glossary and an expanded collection of practice data sets.

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

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

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

Learn all you need to know about seven key innovations disrupting business analytics today. These innovations—the open source business model, cloud analytics, the Hadoop ecosystem, Spark and in-memory analytics, streaming analytics, Deep Learning, and self-service analytics—are radically changing how businesses use data for competitive advantage. Taken together, they are disrupting the business analytics value chain, creating new opportunities. Enterprises who seize the opportunity will thrive and prosper, while others struggle and decline: disrupt or be disrupted. Disruptive Business Analytics provides strategies to profit from disruption. It shows you how to organize for insight, build and provision an open source stack, how to practice lean data warehousing, and how to assimilate disruptive innovations into an organization. Through a short history of business analytics and a detailed survey of products and services, analytics authority Thomas W. Dinsmore provides a practical explanation of the most compelling innovations available today. What You'll Learn Discover how the open source business model works and how to make it work for you See how cloud computing completely changes the economics of analytics Harness the power of Hadoop and its ecosystem Find out why Apache Spark is everywhere Discover the potential of streaming and real-time analytics Learn what Deep Learning can do and why it matters See how self-service analytics can change the way organizations do business Who This Book Is For Corporate actors at all levels of responsibility for analytics: analysts, CIOs, CTOs, strategic decision makers, managers, systems architects, technical marketers, product developers, IT personnel, and consultants.

In the early days of the 20th century, department store magnate JohnWanamaker famously said, "I know that half of my advertising doesn'twork. The problem is that I don't know which half." That remainedbasically true until Google transformed advertising with AdSense basedon new uses of data and analysis. The same might be said about healthcare and it's poised to go through a similar transformation as newtools, techniques, and data sources come on line. Soon we'll makepolicy and resource decisions based on much better understanding ofwhat leads to the best outcomes, and we'll make medical decisionsbased on a patient's specific biology. The result will be betterhealth at less cost. This paper explores how data analysis will help us structure thebusiness of health care more effectively around outcomes, and how itwill transform the practice of medicine by personalizing for eachspecific patient.

Investors and technology gurus have called big data one of the most important trends to come along in decades. Big Data Bootcamp explains what big data is and how you can use it in your company to become one of tomorrow’s market leaders. Along the way, it explains the very latest technologies, companies, and advancements. Big data holds the keys to delivering better customer service, offering more attractive products, and unlocking innovation. That’s why, to remain competitive, every organization should become a big data company. It’s also why every manager and technology professional should become knowledgeable about big data and how it is transforming not just their own industries but the global economy. And that knowledge is just what this book delivers. It explains components of big data like Hadoop and NoSQL databases; how big data is compiled, queried, and analyzed; how to create a big data application; and the business sectors ripe for big data-inspired products and services like retail, healthcare, finance, and education. Best of all, your guide is David Feinleib, renowned entrepreneur, venture capitalist, and author of Why Startups Fail. Feinleib’s Big Data Landscape, a market map featured and explained in the book, is an industry benchmark that has been viewed more than 150,000 times and is used as a reference by VMWare, Dell, Intel, the U.S. Government Accountability Office, and many other organizations. Feinleib also explains: • Why every businessperson needs to understand the fundamentals of big data or get run over by those who do • How big data differs from traditional database management systems • How to create and run a big data project • The technical details powering the big data revolution Whether you’re a Fortune 500 executive or the proprietor of a restaurant or web design studio, Big Data Bootcamp will explain how you can take full advantage of new technologies to transform your company and your career.

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

Much has changed in technology over the past decade. Data is hot, the cloud is ubiquitous, and many organizations need some form of automation. Throughout these transformations, Python has become one of the most popular languages in the world. This practical resource shows you how to use Python for everyday Linux systems administration tasks with today’s most useful DevOps tools, including Docker, Kubernetes, and Terraform. Learning how to interact and automate with Linux is essential for millions of professionals. Python makes it much easier. With this book, you’ll learn how to develop software and solve problems using containers, as well as how to monitor, instrument, load-test, and operationalize your software. Looking for effective ways to "get stuff done" in Python? This is your guide. Python foundations, including a brief introduction to the language How to automate text, write command-line tools, and automate the filesystem Linux utilities, package management, build systems, monitoring and instrumentation, and automated testing Cloud computing, infrastructure as code, Kubernetes, and serverless Machine learning operations and data engineering from a DevOps perspective Building, deploying, and operationalizing a machine learning project

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

Big Data Analytics Using Splunk is a hands-on book showing how to process and derive business value from big data in real time. Examples in the book draw from social media sources such as Twitter (tweets) and Foursquare (check-ins). You also learn to draw from machine data, enabling you to analyze, say, web server log files and patterns of user access in real time, as the access is occurring. Gone are the days when you need be caught out by shifting public opinion or sudden changes in customer behavior. Splunk’s easy to use engine helps you recognize and react in real time, as events are occurring. Splunk is a powerful, yet simple analytical tool fast gaining traction in the fields of big data and operational intelligence. Using Splunk, you can monitor data in real time, or mine your data after the fact. Splunk’s stunning visualizations aid in locating the needle of value in a haystack of a data. Geolocation support spreads your data across a map, allowing you to drill down to geographic areas of interest. Alerts can run in the background and trigger to warn you of shifts or events as they are taking place. With Splunk you can immediately recognize and react to changing trends and shifting public opinion as expressed through social media, and to new patterns of eCommerce and customer behavior. The ability to immediately recognize and react to changing trends provides a tremendous advantage in today’s fast-paced world of Internet business. Big Data Analytics Using Splunk opens the door to an exciting world of real-time operational intelligence. Built around hands-on projects Shows how to mine social media Opens the door to real-time operational intelligence

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

Learn about Cloudera Impala--an open source project that's opening up the Apache Hadoop software stack to a wide audience of database analysts, users, and developers. The Impala massively parallel processing (MPP) engine makes SQL queries of Hadoop data simple enough to be accessible to analysts familiar with SQL and to users of business intelligence tools--and it’s fast enough to be used for interactive exploration and experimentation.

"The chapters in this volume offer useful case studies, technical roadmaps, lessons learned, and a few prescriptions todo this, avoid that.'"-From the Foreword by Joe LaCugna, Ph.D., Enterprise Analytics and Business Intelligence, Starbucks Coffee CompanyWith the growing barrage of "big data," it becomes vitally important for organizations to mak

If you’re a business-minded pioneer who believes that Hadoop can transform your business—and want to learn from the successes and failures of the hearty engineers who have already gone down this trail—Enterprise Hadoop is the guide you've been looking for. Building a Hadoop team, executing a Hadoop project, and even understanding Hadoop constitute perhaps the most daring set of challenges a business leader can take upon himself (or herself) today. Many engineers who tackled Hadoop faced lots of adversity and roadblocks along the way. But that doesn't mean you have to. With this book, you will: Get honest case studies that cut through the hype and provide a realistic look into what it's really like to use Hadoop—good and bad. Understand the intricacies of planning and executing a successful Hadoop project. Learn about the company being interviewed in each case study, including background about their Hadoop team, information on how they started the project, details on how they run their project day-to-day, and items they are considering for the future of their endeavor. Hadoop is not just a program or a tool or a technology, but an ecosystem—a complex collection of tools, projects, techniques, and codebases. If you've previously honed their skills on the client-server, RDBMS, SQL, multi-tiered architectures, Enterprise Hadoopwill guide you through this new landscape. Transforming your business doesn't mean you have to hack your way through the wilderness.

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

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

Data is arriving faster than you can process it and the overall volumes keep growing at a rate that keeps you awake at night. Hadoop can help you tame the data beast. Effective use of Hadoop however requires a mixture of programming, design, and system administration skills. "Hadoop Beginner's Guide" removes the mystery from Hadoop, presenting Hadoop and related technologies with a focus on building working systems and getting the job done, using cloud services to do so when it makes sense. From basic concepts and initial setup through developing applications and keeping the system running as the data grows, the book gives the understanding needed to effectively use Hadoop to solve real world problems. Starting with the basics of installing and configuring Hadoop, the book explains how to develop applications, maintain the system, and how to use additional products to integrate with other systems. While learning different ways to develop applications to run on Hadoop the book also covers tools such as Hive, Sqoop, and Flume that show how Hadoop can be integrated with relational databases and log collection. In addition to examples on Hadoop clusters on Ubuntu uses of cloud services such as Amazon, EC2 and Elastic MapReduce are covered.

There is an ongoing data explosion transpiring that will make previous creations, collections, and storage of data look trivial. Big Data, Mining, and Analytics: Components of Strategic Decision Making ties together big data, data mining, and analytics to explain how readers can leverage them to extract valuable insights from their data. Facilitati

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

Apply cloud design patterns to overcome real-world challenges by building scalable, secure, highly available, and cost-effective solutions Key Features Apply AWS Well-Architected Framework concepts to common real-world use cases Understand how to select AWS patterns and architectures that are best suited to your needs Ensure the security and stability of a solution without impacting cost or performance Book Description One of the most popular cloud platforms in the world, Amazon Web Services (AWS) offers hundreds of services with thousands of features to help you build scalable cloud solutions; however, it can be overwhelming to navigate the vast number of services and decide which ones best suit your requirements. Whether you are an application architect, enterprise architect, developer, or operations engineer, this book will take you through AWS architectural patterns and guide you in selecting the most appropriate services for your projects. AWS for Solutions Architects is a comprehensive guide that covers the essential concepts that you need to know for designing well-architected AWS solutions that solve the challenges organizations face daily. You'll get to grips with AWS architectural principles and patterns by implementing best practices and recommended techniques for real-world use cases. The book will show you how to enhance operational efficiency, security, reliability, performance, and cost-effectiveness using real-world examples. By the end of this AWS book, you'll have gained a clear understanding of how to design AWS architectures using the most appropriate services to meet your organization's technological and business requirements. What you will learn Rationalize the selection of AWS as the right cloud provider for your organization Choose the most appropriate service from AWS for a particular use case or project Implement change and operations management Find out the right resource type and size to balance performance and efficiency Discover how to mitigate risk and enforce security, authentication, and authorization Identify common business scenarios and select the right reference architectures for them Who this book is for This book is for application and enterprise architects, developers, and operations engineers who want to become well-versed with AWS architectural patterns, best practices, and advanced techniques to build scalable, secure, highly available, and cost-effective solutions in the cloud. Although existing AWS users will find this book most useful, it will also help potential users understand how leveraging AWS can benefit their 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 complete instructions for manipulating, processing, cleaning, and crunching datasets in Python. Updated for Python 3.6, the second edition of this hands-on guide is packed with practical case studies that show you how to solve a broad set of data analysis problems effectively. You’ll learn the latest versions of pandas, NumPy, IPython, and Jupyter in the process. Written by Wes McKinney, the creator of the Python pandas project, this book is a practical, modern introduction to data science tools in Python. It’s ideal for analysts new to Python and for Python programmers new to data science and scientific computing. Data files and related material are available on GitHub. Use the IPython shell and Jupyter notebook for exploratory computing Learn basic and advanced features in NumPy (Numerical Python) Get started with data analysis tools in the pandas library Use flexible tools to load, clean, transform, merge, and reshape data Create informative visualizations with matplotlib Apply the pandas groupby facility to slice, dice, and summarize datasets Analyze and manipulate regular and irregular time series data Learn how to solve real-world data analysis problems with thorough, detailed examples

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 Science and Big Data Analytics is about harnessing the power of data for new insights. The book covers the breadth of activities and methods and tools that Data Scientists use. The content focuses on concepts, principles and practical applications that are applicable to any industry and technology environment, and the learning is supported and explained with examples that you can replicate using open-source software. This book will help you: Become a contributor on a data science team Deploy a structured lifecycle approach to data analytics problems Apply appropriate analytic techniques and tools to analyzing big data Learn how to tell a compelling story with data to drive business action Prepare for EMC Proven Professional Data Science Certification Corresponding data sets are available from the book’s page at Wiley which you can find on the Wiley site by searching for the ISBN 9781118876138. Get started discovering, analyzing, visualizing, and presenting data in a meaningful way today!

How do you start? How should you build a plan for cloud migration for your entire portfolio? How will your organization be affected by these changes? This book, based on real-world cloud experiences by enterprise IT teams, seeks to provide the answers to these questions. Here, you’ll see what makes the cloud so compelling to enterprises; with which applications you should start your cloud journey; how your organization will change, and how skill sets will evolve; how to measure progress; how to think about security, compliance, and business buy-in; and how to exploit the ever-growing feature set that the cloud offers to gain strategic and competitive advantage.

Integrating data from multiple sources is essential in the age of big data, but it can be a challenging and time-consuming task. This handy cookbook provides dozens of ready-to-use recipes for using Apache Sqoop, the command-line interface application that optimizes data transfers between relational databases and Hadoop. Sqoop is both powerful and bewildering, but with this cookbook's problem-solution-discussion format, you'll quickly learn how to deploy and then apply Sqoop in your environment. The authors provide MySQL, Oracle, and PostgreSQL database examples on GitHub that you can easily adapt for SQL Server, Netezza, Teradata, or other relational systems. Transfer data from a single database table into your Hadoop ecosystem Keep table data and Hadoop in sync by importing data incrementally Import data from more than one database table Customize transferred data by calling various database functions Export generated, processed, or backed-up data from Hadoop to your database Run Sqoop within Oozie, Hadoop's specialized workflow scheduler Load data into Hadoop's data warehouse (Hive) or database (HBase) Handle installation, connection, and syntax issues common to specific database vendors