For Pro Hadoop By Jason Venner

Dig deep into the data with a hands-on guide to machine learning with updated examples and more! Machine Learning: Hands-On for Developers and Technical Professionals provides hands-on instruction and fully-coded working examples for the most common machine learning techniques used by developers and technical professionals. The book contains a breakdown of each ML variant, explaining how it works and how it is used within certain industries, allowing readers to incorporate the presented techniques into their own work as they follow along. A core tenant of machine learning is a strong focus on data preparation, and a full exploration of the various types of learning algorithms illustrates how the proper tools can help any developer extract information and insights from existing data. The book includes a full complement of Instructor's Materials to facilitate use in the classroom, making this resource useful for students and as a professional reference. At its core, machine learning is a mathematical, algorithm-based technology that forms the basis of historical data mining and modern big data science. Scientific analysis of big data requires a working knowledge of machine learning, which forms predictions based on known properties learned from training data. Machine Learning is an accessible, comprehensive guide for the non-mathematician, providing clear guidance that allows readers to: Learn the languages of machine learning including Hadoop, Mahout, and Weka Understand decision trees, Bayesian networks, and artificial neural networks Implement Association Rule, Real Time, and Batch learning Develop a strategic plan for safe, effective, and efficient machine learning By learning to construct a system that can learn from data, readers can increase their utility across industries. Machine learning sits at the core of deep dive data analysis and visualization, which is increasingly in demand as companies discover the goldmine hiding in their existing data. For the tech professional involved in data science, Machine Learning: Hands-On for Developers and Technical Professionals provides the skills and techniques required to dig deeper.

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.

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.

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

Explains how to use Structured Query Language to work within a relational database system, including information retrieval, security, data manipulation, and user management.

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.

Data Pipelines with Apache Airflow teaches you the ins-and-outs of the Directed Acyclic Graphs (DAGs) that power Airflow, and how to write your own DAGs to meet the needs of your projects. With complete coverage of both foundational and lesser-known features, when you're done you'll be set to start using Airflow for seamless data pipeline development and management. Pipelines can be challenging to manage, especially when your data has to flow through a collection of application components, servers, and cloud services. Airflow lets you schedule, restart, and backfill pipelines, and its easy-to-use UI and workflows with Python scripting has users praising its incredible flexibility. Data Pipelines with Apache Airflow takes you through best practices for creating pipelines for multiple tasks, including data lakes, cloud deployments, and data science. Data Pipelines with Apache Airflow teaches you the ins-and-outs of the Directed Acyclic Graphs (DAGs) that power Airflow, and how to write your own DAGs to meet the needs of your projects. With complete coverage of both foundational and lesser-known features, when you're done you'll be set to start using Airflow for seamless data pipeline development and management. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications.

The new multimedia standards (for example, MPEG-21) facilitate the seamless integration of multiple modalities into interoperable multimedia frameworks, transforming the way people work and interact with multimedia data. These key technologies and multimedia solutions interact and collaborate with each other in increasingly effective ways, contributing to the multimedia revolution and having a significant impact across a wide spectrum of consumer, business, healthcare, education, and governmental domains. This book aims to provide a complete coverage of the areas outlined and to bring together the researchers from academic and industry as well as practitioners to share ideas, challenges, and solutions relating to the multifaceted aspects of this field.

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

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

Pro HadoopApress

Celebrity make-up artist Gary Cockerill is best known for his glamorous, sultry make-up style. Much in demand for his ability to make women look and feel fabulous, he regularly travels the world to get clients ready for red-carpet events, photoshoots and television appearances. In Simply Glamorous, Gary proves that with a little knowledge and practice, everyone can use make-up to enhance or disguise so they feel confident in their appearance. He begins by demonstrating what he loves most about make-up: its ability to transform. By applying four different looks to one face, he shows how the same ‘blank canvas’ can be natural, sophisticated, glamorous or dramatic. He then presents 15 breathtaking makeovers on women of all ages, explaining step by step how to re-create the looks. Divided into ‘Face’, ‘Eyes’ and ‘Lips’, the transformations range from nude and natural – ‘Barely There’, ‘Sun-kissed Goddess’ and ‘Girl Next Door’ – to high octane glamour – ‘Showgirl’, ‘Bombshell’ and ‘Glamour Puss’. They illustrate every aspect of make-up, from contouring, to false eyelashes, to wearing colour, and include the iconic looks that Gary is always asked to create – the smoky eye, the fifties flick and the red lip. ‘Tutorials’ address key topics in greater detail, such as skincare, foundation, contouring and application techniques for eyes and lips. As a self-taught make-up artist, Gary works instinctively with amazing results. His approach is creative and painterly, and his teaching style is friendly and down-to-earth as he shows how glamorous make-up can be accessible to everyone. He may not always do things the conventional way, but his techniques are the result of more than 25 years spent working with women and men of all ages, ethnicities and styles. He knows what works and what looks good, and encourages everyone to have fun and experiment with make-up to make themselves look and feel beautiful and glamorous.

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.

The ASP.NET MVC 3 Framework is the latest evolution of Microsoft’s ASP.NET web platform. It provides a high-productivity programming model that promotes cleaner code architecture, test-driven development, and powerful extensibility, combined with all the benefits of ASP.NET 4. In this third edition, the core model-view-controller (MVC) architectural concepts are not simply explained or discussed in

isolation, but are demonstrated in action. You'll work through an extended tutorial to create a working e-commerce web application that combines ASP.NET MVC with the latest C# language features and unit-testing best practices. By gaining this invaluable, practical experience, you'll discover MVC's strengths and weaknesses for yourself—and put your best-learned theory into practice. The book's authors Steve Sanderson and Adam Freeman have both watched the growth of ASP.NET MVC since its first release. Steve is a well-known blogger on the MVC Framework and a member of the Microsoft Web Platform and Tools team. Adam started designing and building web applications 15 years ago and has been responsible for some of the world's largest and most ambitious projects. You can be sure you are in safe hands.

Learn how to take full advantage of Apache Kafka, the distributed, publish-subscribe queue for handling real-time data feeds. With this comprehensive book, you will understand how Kafka works and how it is designed. Authors Neha Narkhede, Gwen Shapira, and Todd Palino show you how to deploy production Kafka clusters; secure, tune, and monitor them; write rock-solid applications that use Kafka; and build scalable stream-processing applications. Learn how Kafka compares to other queues, and where it fits in the big data ecosystem. Dive into Kafka's internal designPick up best practices for developing applications that use Kafka. Understand the best way to deploy Kafka in production monitoring, tuning, and maintenance tasks. Learn how to secure a Kafka cluster.

Summary Big Data teaches you to build big data systems using an architecture that takes advantage of clustered hardware along with new tools designed specifically to capture and analyze web-scale data. It describes a scalable, easy-to-understand approach to big data systems that can be built and run by a small team. Following a realistic example, this book guides readers through the theory of big data systems, how to implement them in practice, and how to deploy and operate them once they're built. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Book Web-scale applications like social networks, real-time analytics, or e-commerce sites deal with a lot of data, whose volume and velocity exceed the limits of traditional database systems. These applications require architectures built around clusters of machines to store and process data of any size, or speed. Fortunately, scale and simplicity are not mutually exclusive. Big Data teaches you to build big data systems using an architecture designed specifically to capture and analyze web-scale data. This book presents the Lambda Architecture, a scalable, easy-to-understand approach that can be built and run by a small team. You'll explore the theory of big data systems and how to implement them in practice. In addition to discovering a general framework for processing big data, you'll learn specific technologies like Hadoop, Storm, and NoSQL databases. This book requires no previous exposure to large-scale data analysis or NoSQL tools. Familiarity with traditional databases is helpful. What's Inside Introduction to big data systems Real-time processing of web-scale data Tools like Hadoop, Cassandra, and Storm Extensions to traditional database skills About the Authors Nathan Marz is the creator of Apache Storm and the originator of the Lambda Architecture for big data systems. James Warren is an analytics architect with a background in machine learning and scientific computing. Table of Contents A new paradigm for Big Data PART 1 BATCH LAYER Data model for Big Data Data model for Big Data: Illustration Data storage on the batch layer Data storage on the batch layer: Illustration Batch layer Batch layer: Illustration An example batch layer: Architecture and algorithms An example batch layer: Implementation PART 2 SERVING LAYER Serving layer Serving layer: Illustration PART 3 SPEED LAYER Realtime views Realtime views: Illustration Queuing and stream processing Queuing and stream processing: Illustration Micro-batch stream processing Micro-batch stream processing: Illustration Lambda Architecture in depth

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

This book provides readers the "big picture" and a comprehensive survey of the domain of big data processing systems. For the past decade, the Hadoop framework has dominated the world of big data processing, yet recently academia and industry have started to recognize its limitations in several application domains and big data processing scenarios such as the large-scale processing of structured data, graph data and streaming data. Thus, it is now gradually being replaced by a collection of engines that are dedicated to specific verticals (e.g. structured data, graph data, and streaming data). The book explores this new wave of systems, which it refers to as Big Data 2.0 processing systems. After Chapter 1 presents the general background of the big data phenomena, Chapter 2 provides an overview of various general-purpose big data processing systems that allow their users to develop various big data processing jobs for different application domains. In turn, Chapter 3 examines various systems that have been introduced to support the SQL flavor on top of the Hadoop infrastructure and provide competing and scalable performance in the processing of large-scale structured data. Chapter 4 discusses several systems that have been designed to tackle the problem of large-scale graph processing, while the main focus of Chapter 5 is on several systems that have been designed to provide scalable solutions for processing big data streams, and on other sets of systems that have been introduced to support the development of data pipelines between various types of big data processing jobs and systems. Lastly, Chapter 6 shares conclusions and an outlook on future research challenges. Overall, the book offers a valuable reference guide for students, researchers and professionals in the domain of big data processing systems. Further, its comprehensive content will hopefully encourage readers to pursue further research on the subject.

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.

Practical Ext JS 4 will get you up and running, using Ext JS 4.2 for your projects, as quickly as possible. After a quick refresher on some JavaScript basics, you will get to grips with Ext JS 4's OO concepts (such as mixins) and familiarize yourself with its UI components and layout. You'll learn all the core features of the Ext JS framework, such as its MVC architecture, theming and styling your applications, and displaying data through components such as grids, trees, and charts. You'll use the Ext JS components and create an entire application from scratch by following the many practical examples. Finally, you'll learn about unit testing and packaging to build and deploy better applications. Provides you with a solid knowledge of the building blocks of Ext JS 4 Takes you through developing applications using the MVC architecture Demonstrates extending the UI with custom components and plugins Shows you how to unit test Ext JS 4 applications with Jasmine and deploy them with Sencha Cmd Completely up-to-date for the latest Ext JS 4.2

Based on Big Nerd Ranch's popular iPhone Bootcamp class, iPhone Programming: The Big Nerd Ranch Guide leads you through the essential tools and techniques for developing applications for the iPhone, iPad, and iPod Touch. In each chapter, you will learn programming concepts and apply them immediately as you build an application or enhance one from a previous chapter. These applications have been carefully designed and tested to teach the associated concepts and to provide practice working with the standard development tools Xcode, Interface Builder, and Instruments. The guide's learn-while-doing approach delivers the practical knowledge and experience you need to design and build real-world applications. Here are some of the topics covered: Dynamic interfaces with animation Using the camera and photo library User location and mapping services Accessing accelerometer data Handling multi-touch gestures Navigation and tabbed applications Tables and creating custom rows Multiple ways of storing and loading data: archiving, Core Data, SQLite Communicating with web services ALocalization/Internationalization "After many 'false starts' with other iPhone development books, these clear and concise tutorials made the concepts gel for me. This book is a definite must have for any budding iPhone developer." –Peter Watling, New Zealand, Developer of BubbleWrap

This IBM® Redbooks® publication highlights IBM Technical Computing as a flexible infrastructure for clients looking to reduce capital and operational expenditures, optimize energy usage, or re-use the infrastructure. This book strengthens IBM SmartCloud® solutions, in particular IBM Technical Computing clouds, with a well-defined and documented deployment model within an IBM System x® or an IBM Flex SystemTM. This provides clients with a cost-effective, highly scalable, robust solution with a planned foundation for scaling, capacity, resilience, optimization, automation, and monitoring. This book is targeted toward technical professionals (consultants, technical support staff, IT Architects, and IT Specialists) responsible for providing cloud-computing solutions and support.

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.

ICSSCCET 2015 will be the most comprehensive conference focused on the various aspects of advances in Systems, Science, Management, Medical Sciences, Communication, Engineering, Technology, Interdisciplinary Research Theory and Technology. This Conference provides a chance for academic and industry professionals to discuss recent progress in the area of Interdisciplinary Research Theory and Technology. Furthermore, we expect that the conference and its publications will be a trigger for further related research and technology improvements in this important subject. The goal of this conference is to bring together the researchers from academia and industry as well as practitioners to share ideas, problems and solutions relating to the multifaceted aspects of Interdisciplinary Research Theory and Technology.

Gain a broad foundation of advanced data analytics concepts and discover the recent revolution in databases such as Neo4j, Elasticsearch, and MongoDB. This book discusses how to implement ETL

techniques including topical crawling, which is applied in domains such as high-frequency algorithmic trading and goal-oriented dialog systems. You'll also see examples of machine learning concepts such as semi-supervised learning, deep learning, and NLP. Advanced Data Analytics Using Python also covers important traditional data analysis techniques such as time series and principal component analysis. After reading this book you will have experience of every technical aspect of an analytics project. You'll get to know the concepts using Python code, giving you samples to use in your own projects. What You Will Learn Work with data analysis techniques such as classification, clustering, regression, and forecasting Handle structured and unstructured data, ETL techniques, and different kinds of databases such as Neo4j, Elasticsearch, MongoDB, and MySQL Examine the different big data frameworks, including Hadoop and Spark Discover advanced machine learning concepts such as semi-supervised learning, deep learning, and NLP Who This Book Is For Data scientists and software developers interested in the field of data analytics.

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.

Pro SQL Server 2012 Practices is an anthology of high-end wisdom from a group of accomplished database administrators who are quietly but relentlessly pushing the performance and feature envelope of Microsoft SQL Server 2012. With an emphasis upon performance—but also branching into release management, auditing, and other issues—the book helps you deliver the most value for your company's investment in Microsoft's flagship database system. Goes beyond the manual to cover good techniques and best practices Delivers knowledge usually gained only by hard experience Focuses upon performance, scalability, reliability Helps achieve the predictability needed to be in control at all times

Navy analysts are struggling to keep pace with the growing flood of data collected by intelligence, surveillance, and reconnaissance sensors. This challenge is sure to intensify as the Navy continues to field new and additional sensors. The authors explore options for solving the Navy's "big data" challenge, considering changes across four dimensions: people, tools and technology, data and data architectures, and demand and demand management.

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

Over a half-million sold! The sequel, The Unicorn Project, is coming Nov 26 "Every person involved in a failed IT project should be forced to read this book."—TIM O'REILLY, Founder & CEO of O'Reilly Media "The Phoenix Project is a must read for business and IT executives who are struggling with the growing complexity of IT."—JIM WHITEHURST, President and CEO, Red Hat, Inc. Five years after this sleeper hit took on the world of IT and flipped it on it's head, the 5th Anniversary Edition of The Phoenix Project continues to guide IT in the DevOps revolution. In this newly updated and expanded edition of the bestselling The Phoenix Project, co-author Gene Kim includes a new afterword and a deeper delve into the Three Ways as described in The DevOps Handbook. Bill, an IT manager at Parts Unlimited, has been tasked with taking on a project critical to the future of the business, code named Phoenix Project. But the project is massively over budget and behind schedule. The CEO demands Bill must fix the mess in ninety days or else Bill's entire department will be outsourced. With the help of a prospective board member and his mysterious philosophy of The Three Ways, Bill starts to see that IT work has more in common with a manufacturing plant work than he ever imagined. With the clock ticking, Bill must organize work flow streamline interdepartmental communications, and effectively serve the other business functions at Parts Unlimited. In a fast-paced and entertaining style, three luminaries of the DevOps movement deliver a story that anyone who works in IT will recognize. Readers will not only learn how to improve their own IT organizations, they'll never view IT the same way again. "This book is a gripping read that captures brilliantly the dilemmas that face companies which depend on IT, and offers real-world solutions."—JEZ HUMBLE, Co-author of Continuous Delivery, Lean Enterprise, Accelerate, and The DevOps Handbook ———— "I'm delighted at how The Phoenix Project has reshaped so many conversations in technology. My goal in writing The Unicorn Project was to explore and reveal the necessary but invisible structures required to make developers (and all engineers) productive, and reveal the devastating effects of technical debt and complexity. I hope this book can create common ground for technology and business leaders to leave the past behind, and co-create a better future together."—Gene Kim, November 2019

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

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 indespensible for anyone using Hadoop." --Tom White, author of Hadoop: The

Definitive Guide

Class-tested and coherent, this textbook teaches classical and web information retrieval, including web search and the related areas of text classification and text clustering from basic concepts. It gives an up-to-date treatment of all aspects of the design and implementation of systems for gathering, indexing, and searching documents; methods for evaluating systems; and an introduction to the use of machine learning methods on text collections. All the important ideas are explained using examples and figures, making it perfect for introductory courses in information retrieval for advanced undergraduates and graduate students in computer science. Based on feedback from extensive classroom experience, the book has been carefully structured in order to make teaching more natural and effective. Slides and additional exercises (with solutions for lecturers) are also available through the book's supporting website to help course instructors prepare their lectures.

Tap into the wisdom of experts to learn what every programmer should know, no matter what language you use. With the 97 short and extremely useful tips for programmers in this book, you'll expand your skills by adopting new approaches to old problems, learning appropriate best practices, and honing your craft through sound advice. With contributions from some of the most experienced and respected practitioners in the industry--including Michael Feathers, Pete Goodliffe, Diomidis Spinellis, Cay Horstmann, Verity Stob, and many more--this book contains practical knowledge and principles that you can apply to all kinds of projects. A few of the 97 things you should know: "Code in the Language of the Domain" by Dan North "Write Tests for People" by Gerard Meszaros "Convenience Is Not an -ility" by Gregor Hohpe "Know Your IDE" by Heinz Kabutz "A Message to the Future" by Linda Rising "The Boy Scout Rule" by Robert C. Martin (Uncle Bob) "Beware the Share" by Udi Dahan

Designing Cloud Data Platforms teaches you how to integrate data from multiple sources into a single, cloud-based, modern data platform. In it, you'll follow a six-layer approach to creating cloud data platforms that maximizes flexibility and manageability and reduces costs. Companies that embrace modern cloud data platforms benefit from an integrated view of their business using all of their data and can take advantage of advanced analytic practices to drive predictions and as yet unimagined data services. Designing Cloud Data Platforms is a hands-on guide to envisioning and designing a modern scalable data platform that takes full advantage of the flexibility of the cloud. Designing Cloud Data Platforms teaches you how to integrate data from multiple sources into a single, cloud-based, modern data platform. In it, you'll follow a six-layer approach to creating cloud data platforms that maximizes flexibility and manageability and reduces costs. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications.

Data science libraries, frameworks, modules, and toolkits are great for doing data science, but they’re also a good way to dive into the discipline without actually understanding data science. In this book, you’ll learn how many of the most fundamental data science tools and algorithms work by implementing them from scratch. If you have an aptitude for mathematics and some programming skills, author Joel Grus will help you get comfortable with the math and statistics at the core of data science, and with hacking skills you need to get started as a data scientist. Today’s messy glut of data holds answers to questions no one’s even thought to ask. This book provides you with the know-how to dig those answers out. Get a crash course in Python Learn the basics of linear algebra, statistics, and probability—and understand how and when they're used in data science Collect, explore, clean, munge, and manipulate data Dive into the fundamentals of machine learning Implement models such as k-nearest Neighbors, Naive Bayes, linear and logistic regression, decision trees, neural networks, and clustering Explore recommender systems, natural language processing, network analysis, MapReduce, and databases

Want to tap the power behind search rankings, product recommendations, social bookmarking, and online matchmaking? This fascinating book demonstrates how you can build Web 2.0 applications to mine the enormous amount of data created by people on the Internet. With the sophisticated algorithms in this book, you can write smart programs to access interesting datasets from other web sites, collect data from users of your own applications, and analyze and understand the data once you've found it. Programming Collective Intelligence takes you into the world of machine learning and statistics, and explains how to draw conclusions about user experience, marketing, personal tastes, and human behavior in general -- all from information that you and others collect every day. Each algorithm is described clearly and concisely with code that can immediately be used on your web site, blog, Wiki, or specialized application. This book explains: Collaborative filtering techniques that enable online retailers to recommend products or media Methods of clustering to detect groups of similar items in a large dataset Search engine features -- crawlers, indexers, query engines, and the PageRank algorithm Optimization algorithms that search millions of possible solutions to a problem and choose the best one Bayesian filtering, used in spam filters for classifying documents based on word types and other features Using decision trees not only to make predictions, but to model the way decisions are made Predicting numerical values rather than classifications to build price models Support vector machines to match people in online dating sites Non-negative matrix factorization to find the independent features in a dataset Evolving intelligence for problem solving -- how a computer develops its skill by improving its own code the more it plays a game Each chapter includes exercises for extending the algorithms to make them more powerful. Go beyond simple database-backed applications and put the wealth of Internet data to work for you. "Bravo! I cannot think of a better way for a developer to first learn these algorithms and methods, nor can I think of a better way for me (an old AI dog) to reinvigorate my knowledge of the details." -- Dan Russell, Google "Toby's book does a great job of breaking down the complex subject matter of machine-learning algorithms into practical, easy-to-understand examples that can be directly applied to analysis of social interaction across the Web today. If I had this book two years ago, it would have saved precious time going down some fruitless paths." -- Tim Wolters, CTO, Collective Intellect

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

Written by renowned data science experts Foster Provost and Tom Fawcett, Data Science for Business introduces the fundamental principles of data science, and walks you through the "data-analytic thinking" necessary for extracting useful knowledge and business value from the data you collect. This guide also helps you understand the many data-mining techniques in use today. Based on an MBA course Provost has taught at New York University over the past ten years, Data Science for Business provides examples of real-world business problems to illustrate these principles. You'll not only learn how to improve communication between business stakeholders and data scientists, but also how participate intelligently in your company's data science projects. You'll also discover how to think data-analytically, and fully appreciate how data science methods can support business decision-making. Understand how data science fits in your organization—and how you can use it for competitive advantage Treat data as a business asset that requires careful investment if you're to gain real value Approach business problems data-analytically, using the data-mining process to gather good data in the most appropriate way Learn general concepts for actually extracting knowledge from data Apply data science principles when interviewing data science job candidates

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