

Practical Hadoop By Example Nyoug

Big data solutions enable us to change how we do business by exploiting previously unused sources of information in ways that were not possible just a few years ago. In IBM® Smarter Planet® terms, big data helps us to change the way that the world works. The purpose of this IBM Redpaper™ publication is to consider the performance and capacity implications of big data solutions, which must be taken into account for them to be viable. This paper describes the benefits that big data approaches can provide. We then cover performance and capacity considerations for creating big data solutions. We conclude with what this means for big data solutions, both now and in the future. Intended readers for this paper include decision-makers, consultants, and IT architects.

This invaluable addition to any data scientist's library shows you how to apply the R programming language and useful statistical techniques to everyday business situations as well as how to effectively present results to audiences of all levels. To answer the ever-increasing demand for machine learning and analysis, this new edition boasts additional R tools, modeling techniques, and more. Practical Data Science with R, Second Edition takes a practice-oriented approach to explaining basic principles in the ever-expanding field of data science. You'll jump right to real-world use cases as you apply the R programming language and statistical analysis techniques to carefully explained examples based in marketing, business intelligence, and decision support. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications.

The world contains an unimaginably vast amount of digital information which is getting ever vaster ever more rapidly. This makes it possible to do many things that previously could not be done: spot business trends, prevent diseases, combat crime and so on. Managed well, the textual data can be used to unlock new sources of economic value, provide fresh insights into science and hold governments to account. As the Internet expands and our natural capacity to process the unstructured text that it contains diminishes, the value of text mining for information retrieval and search will increase dramatically. This comprehensive professional reference brings together all the information, tools and methods a professional will need to efficiently use text mining applications and statistical analysis. The Handbook of Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications presents a comprehensive how-to reference that shows the user how to conduct text mining and statistically analyze results. In addition to providing an in-depth examination of core text mining and link detection tools, methods and operations, the book examines advanced preprocessing techniques, knowledge representation considerations, and visualization approaches. Finally, the book explores current real-world, mission-critical applications of text mining and link detection using real world example tutorials in such varied fields as corporate, finance, business

intelligence, genomics research, and counterterrorism activities. -Extensive case studies, most in a tutorial format, allow the reader to 'click through' the example using a software program, thus learning to conduct text mining analyses in the most rapid manner of learning possible -Numerous examples, tutorials, power points and datasets available via companion website on Elsevierdirect.com -Glossary of text mining terms provided in the appendix

This volume comprises the select proceedings of the annual convention of the Computer Society of India. Divided into 10 topical volumes, the proceedings present papers on state-of-the-art research, surveys, and succinct reviews. The volumes cover diverse topics ranging from communications networks to big data analytics, and from system architecture to cyber security. This volume focuses on Big Data Analytics. The contents of this book will be useful to researchers and students alike.

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

The best-selling author of Big Data is back, this time with a unique and in-depth insight into how specific companies use big data. Big data is on the tip of everyone's tongue. Everyone understands its power and importance, but many fail to grasp the actionable steps and resources required to utilise it effectively. This book fills the knowledge gap by showing how major companies are using big data every day, from an up-close, on-the-ground perspective. From technology, media and retail, to sport teams, government agencies and financial institutions, learn the actual strategies and processes being used to learn about customers, improve manufacturing, spur innovation, improve safety and so much more.

Organised for easy dip-in navigation, each chapter follows the same structure to give you the information you need quickly. For each company profiled, learn what data was used, what problem it solved and the processes put it place to make it practical, as well as the technical details, challenges and lessons learned from

each unique scenario. Learn how predictive analytics helps Amazon, Target, John Deere and Apple understand their customers Discover how big data is behind the success of Walmart, LinkedIn, Microsoft and more Learn how big data is changing medicine, law enforcement, hospitality, fashion, science and banking Develop your own big data strategy by accessing additional reading materials at the end of each chapter

Dive into the world of SQL on Hadoop and get the most out of your Hive data warehouses. This book is your go-to resource for using Hive: authors Scott Shaw, Ankur Gupta, David Kjerrumgaard, and Andreas Francois Vermeulen take you through learning HiveQL, the SQL-like language specific to Hive, to analyze, export, and massage the data stored across your Hadoop environment. From deploying Hive on your hardware or virtual machine and setting up its initial configuration to learning how Hive interacts with Hadoop, MapReduce, Tez and other big data technologies, Practical Hive gives you a detailed treatment of the software. In addition, this book discusses the value of open source software, Hive performance tuning, and how to leverage semi-structured and unstructured data. What You Will Learn Install and configure Hive for new and existing datasets Perform DDL operations Execute efficient DML operations Use tables, partitions, buckets, and user-defined functions Discover performance tuning tips and Hive best practices Who This Book Is For Developers, companies, and professionals who deal with large amounts of data and could use software that can efficiently manage large volumes of input. It is assumed that readers have the ability to work with SQL.

Filled with practical, step-by-step instructions and clear explanations for the most important and useful tasks. Instant Apache Sqoop is full of step-by-step instructions and practical examples along with challenges to test and improve your knowledge. This book is great for developers who are looking to get a good grounding in how to effectively and efficiently move data between RDBMS and the Hadoop ecosystem. It's assumed that you will have some experience in Hadoop already as well as some familiarity with HBase and Hive.

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

Mllib and productionize models using MLflow

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

This is the eBook of the printed book and may not include any media, website access codes, or print supplements that may come packaged with the bound book. The Comprehensive, Up-to-Date Apache Hadoop Administration Handbook and Reference “Sam Alapati has worked with production Hadoop clusters for six years. His unique depth of experience has enabled him to write the go-to resource for all administrators looking to spec, size, expand, and secure production Hadoop clusters of any size.” —Paul Dix, Series Editor In Expert Hadoop® Administration, leading Hadoop administrator Sam R. Alapati brings together authoritative knowledge for creating, configuring, securing, managing, and optimizing production Hadoop clusters in any environment. Drawing on his experience with large-scale Hadoop administration, Alapati integrates action-oriented advice with carefully researched explanations of both problems and solutions. He covers an unmatched range of topics and offers an unparalleled collection of realistic examples. Alapati demystifies complex Hadoop environments, helping you understand exactly what happens behind the scenes when you administer your cluster. You’ll gain unprecedented insight as you walk through building clusters from scratch and configuring high availability, performance, security, encryption, and other key attributes. The high-value administration skills you learn here will be indispensable no matter what Hadoop distribution you use or what Hadoop applications you run. Understand Hadoop’s architecture from an administrator’s standpoint Create simple and fully distributed clusters Run MapReduce and Spark applications in a Hadoop cluster Manage and protect Hadoop data and high availability Work with HDFS commands, file permissions, and storage management Move data, and use YARN to allocate resources and schedule jobs Manage job workflows with Oozie and Hue Secure, monitor, log, and optimize Hadoop Benchmark and troubleshoot Hadoop

Big Data represents a new era in data exploration and utilization, and IBM is uniquely positioned to help clients navigate this transformation. This book reveals how IBM is leveraging open source Big Data technology, infused with IBM technologies, to deliver a robust, secure, highly available, enterprise-class Big Data platform. The three defining characteristics of Big Data--volume, variety, and velocity--are discussed. You'll get a primer on Hadoop and how IBM is hardening it for the enterprise, and learn when to leverage IBM InfoSphere BigInsights (Big Data at rest) and IBM InfoSphere Streams (Big Data in motion) technologies. Industry use cases are also included in this practical guide. Learn how IBM hardens Hadoop for enterprise-class scalability and reliability Gain insight into IBM's unique in-motion and at-rest Big Data analytics platform Learn tips and tricks for Big Data use cases and solutions Get a quick Hadoop primer Re-architect relational applications to NoSQL, integrate relational database

management systems with the Hadoop ecosystem, and transform and migrate relational data to and from Hadoop components. This book covers the best-practice design approaches to re-architecting your relational applications and transforming your relational data to optimize concurrency, security, denormalization, and performance. Winner of IBM's 2012 Gerstner Award for his implementation of big data and data warehouse initiatives and author of Practical Hadoop Security, author Bhushan Lakhe walks you through the entire transition process. First, he lays out the criteria for deciding what blend of re-architecting, migration, and integration between RDBMS and HDFS best meets your transition objectives. Then he demonstrates how to design your transition model. Lakhe proceeds to cover the selection criteria for ETL tools, the implementation steps for migration with SQOOP- and Flume-based data transfers, and transition optimization techniques for tuning partitions, scheduling aggregations, and redesigning ETL. Finally, he assesses the pros and cons of data lakes and Lambda architecture as integrative solutions and illustrates their implementation with real-world case studies. Hadoop/NoSQL solutions do not offer by default certain relational technology features such as role-based access control, locking for concurrent updates, and various tools for measuring and enhancing performance. Practical Hadoop Migration shows how to use open-source tools to emulate such relational functionalities in Hadoop ecosystem components. What You'll Learn Decide whether you should migrate your relational applications to big data technologies or integrate them Transition your relational applications to Hadoop/NoSQL platforms in terms of logical design and physical implementation Discover RDBMS-to-HDFS integration, data transformation, and optimization techniques Consider when to use Lambda architecture and data lake solutions Select and implement Hadoop-based components and applications to speed transition, optimize integrated performance, and emulate relational functionalities Who This Book Is For Database developers, database administrators, enterprise architects, Hadoop/NoSQL developers, and IT leaders. Its secondary readership is project and program managers and advanced students of database and management information systems.

Summary Hadoop in Practice, Second Edition provides over 100 tested, instantly useful techniques that will help you conquer big data, using Hadoop. This revised new edition covers changes and new features in the Hadoop core architecture, including MapReduce 2. Brand new chapters cover YARN and integrating Kafka, Impala, and Spark SQL with Hadoop. You'll also get new and updated techniques for Flume, Sqoop, and Mahout, all of which have seen major new versions recently. In short, this is the most practical, up-to-date coverage of Hadoop available anywhere. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Book It's always a good time to upgrade your Hadoop skills! Hadoop in Practice, Second Edition provides a collection of 104 tested, instantly useful techniques for analyzing real-time streams, moving data securely, machine learning, managing

large-scale clusters, and taming big data using Hadoop. This completely revised edition covers changes and new features in Hadoop core, including MapReduce 2 and YARN. You'll pick up hands-on best practices for integrating Spark, Kafka, and Impala with Hadoop, and get new and updated techniques for the latest versions of Flume, Sqoop, and Mahout. In short, this is the most practical, up-to-date coverage of Hadoop available. Readers need to know a programming language like Java and have basic familiarity with Hadoop. What's Inside Thoroughly updated for Hadoop 2 How to write YARN applications Integrate real-time technologies like Storm, Impala, and Spark Predictive analytics using Mahout and RR Readers need to know a programming language like Java and have basic familiarity with Hadoop. About the Author Alex Holmes works on tough big-data problems. He is a software engineer, author, speaker, and blogger specializing in large-scale Hadoop projects. Table of Contents PART 1 BACKGROUND AND FUNDAMENTALS Hadoop in a heartbeat Introduction to YARN PART 2 DATA LOGISTICS Data serialization—working with text and beyond Organizing and optimizing data in HDFS Moving data into and out of Hadoop PART 3 BIG DATA PATTERNS Applying MapReduce patterns to big data Utilizing data structures and algorithms at scale Tuning, debugging, and testing PART 4 BEYOND MAPREDUCE SQL on Hadoop Writing a YARN application

Go ahead, be skeptical about big data. The author was—at first. When the term “big data” first came on the scene, bestselling author Tom Davenport (Competing on Analytics, Analytics at Work) thought it was just another example of technology hype. But his research in the years that followed changed his mind. Now, in clear, conversational language, Davenport explains what big data means—and why everyone in business needs to know about it. Big Data at Work covers all the bases: what big data means from a technical, consumer, and management perspective; what its opportunities and costs are; where it can have real business impact; and which aspects of this hot topic have been oversold. This book will help you understand:

- Why big data is important to you and your organization
- What technology you need to manage it
- How big data could change your job, your company, and your industry
- How to hire, rent, or develop the kinds of people who make big data work
- The key success factors in implementing any big data project
- How big data is leading to a new approach to managing analytics

With dozens of company examples, including UPS, GE, Amazon, United Healthcare, Citigroup, and many others, this book will help you seize all opportunities—from improving decisions, products, and services to strengthening customer relationships. It will show you how to put big data to work in your own organization so that you too can harness the power of this ever-evolving new resource.

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 to 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.

Audience and media analytics is more important now than ever, and this latest volume in the cutting-edge BEA Electronic Media Research Series collects some of the top scholars working with big data and analytics today. These chapters describe the development and help define media analytics as an academic discipline and professional practice. Understanding audiences is integral to creating and distributing media messages and the study of media analytics requires knowing a range of skills including research methods, the necessary tools available, familiarity with statistical procedures, and a mindset to provide insights and apply findings. This book summarizes the insights of analytics practitioners regarding the current state of legacy media analysis and social media analytics. Topics covered include the evolution of media technologies, the teaching of media measurement and analytics, the transition taking place in media research, and the use of media analytics to answer meaningful questions,

drive content creation, and engage with audiences.

Although interest in machine learning has reached a high point, lofty expectations often scuttle projects before they get very far. How can machine learning—especially deep neural networks—make a real difference in your organization? This hands-on guide not only provides the most practical information available on the subject, but also helps you get started building efficient deep learning networks. Authors Adam Gibson and Josh Patterson provide theory on deep learning before introducing their open-source Deeplearning4j (DL4J) library for developing production-class workflows. Through real-world examples, you'll learn methods and strategies for training deep network architectures and running deep learning workflows on Spark and Hadoop with DL4J. Dive into machine learning concepts in general, as well as deep learning in particular Understand how deep networks evolved from neural network fundamentals Explore the major deep network architectures, including Convolutional and Recurrent Learn how to map specific deep networks to the right problem Walk through the fundamentals of tuning general neural networks and specific deep network architectures Use vectorization techniques for different data types with DataVec, DL4J's workflow tool Learn how to use DL4J natively on Spark and Hadoop

The world of medical technologies is undergoing a sea change in the domain of consumer culture. Having a grasp on what appeals to consumers and how consumers are making purchasing decisions is essential to the success of any organization that thrives by offering a product or service. As such, it is vital to examine the consumer-centered aspects of medical technological developments that have a patient-centered focus and allow patients to take part in their own personal health and wellness. Consumer-Driven Technologies in Healthcare: Breakthroughs in Research and Practice is a critical source of academic knowledge on the use of smartphones and other technological devices for cancer therapy, fitness and wellness, chronic disease monitoring, and other areas. The tracking of these items using technology has allowed consumers to take control of their own healthcare. Highlighting a range of pertinent topics such as clinical decision support systems, patient engagement, and electronic health records, this publication is an ideal reference source for doctors, nurse practitioners, hospital administrators, medical professionals, IT professionals, academicians, and researchers interested in advancing medical practice through technology.

Building models is a small part of the story when it comes to deploying machine learning applications. The entire process involves developing, orchestrating, deploying, and running scalable and portable machine learning workloads--a process Kubeflow makes much easier. This practical book shows data scientists, data engineers, and platform architects how to plan and execute a Kubeflow project to make their Kubernetes workflows portable and scalable. Authors Josh Patterson, Michael Katzenellenbogen, and Austin Harris demonstrate how this open source platform orchestrates workflows by managing machine learning pipelines. You'll learn how to plan and execute a Kubeflow platform that can support workflows from on-premises to cloud providers including Google, Amazon, and Microsoft. Dive into Kubeflow architecture and learn best practices for using the platform Understand the process of planning your Kubeflow deployment Install Kubeflow on an existing on-premises Kubernetes cluster Deploy Kubeflow on Google Cloud Platform step-by-step from the

command line Use the managed Amazon Elastic Kubernetes Service (EKS) to deploy Kubeflow on AWS Deploy and manage Kubeflow across a network of Azure cloud data centers around the world Use KFServing to develop and deploy machine learning models

Summary Mahout in Action is a hands-on introduction to machine learning with Apache Mahout. Following real-world examples, the book presents practical use cases and then illustrates how Mahout can be applied to solve them. Includes a free audio- and video-enhanced ebook. About the Technology A computer system that learns and adapts as it collects data can be really powerful. Mahout, Apache's open source machine learning project, captures the core algorithms of recommendation systems, classification, and clustering in ready-to-use, scalable libraries. With Mahout, you can immediately apply to your own projects the machine learning techniques that drive Amazon, Netflix, and others. About this Book This book covers machine learning using Apache Mahout.

Based on experience with real-world applications, it introduces practical use cases and illustrates how Mahout can be applied to solve them. It places particular focus on issues of scalability and how to apply these techniques against large data sets using the Apache Hadoop framework. This book is written for developers familiar with Java -- no prior experience with Mahout is assumed. Owners of a Manning pBook purchased anywhere in the world can download a free eBook from manning.com at any time. They can do so multiple times and in any or all formats available (PDF, ePub or Kindle). To do so, customers must register their printed copy on Manning's site by creating a user account and then following instructions printed on the pBook registration insert at the front of the book. What's Inside Use group data to make individual recommendations Find logical clusters within your data Filter and refine with on-the-fly classification Free audio and video extras Table of Contents Meet Apache Mahout PART 1

RECOMMENDATIONS Introducing recommenders Representing recommender data Making recommendations Taking recommenders to production Distributing recommendation computations PART 2 CLUSTERING Introduction to clustering Representing data Clustering algorithms in Mahout Evaluating and improving clustering quality Taking clustering to production Real-world applications of clustering PART 3 CLASSIFICATION Introduction to classification Training a classifier Evaluating and tuning a classifier Deploying a classifier Case study: Shop It To Me

This book is useful for Hadoop administrators who need to learn how to monitor and diagnose their clusters. Also, the book will prove useful for new users of the technology, as the language used is simple and easy to grasp.

Construct a robust end-to-end solution for analyzing and visualizing streaming data Real-time analytics is the hottest topic in data analyticstoday. In Real-Time Analytics: Techniques to Analyze and Visualize Streaming Data, expert Byron Ellis teaches dataanalysts technologies to build an effective real-time analyticsplatform. This platform can then be used to make sense of theconstantly changing data that is beginning to outpace traditionalbatch-based analysis platforms. The author is among a very few leading experts in the field. Hehas a prestigious background in research, development, analytics,real-time visualization, and Big Data streaming and is uniquelyqualified to help you explore this revolutionary field. Moving froma description of the overall analytic architecture of real-timeanalytics to using specific tools to obtain targeted results,Real-Time Analytics leverages open source and moderncommercial tools to construct robust,

efficient systems that can provide real-time analysis in a cost-effective manner. The book includes: A deep discussion of streaming data systems and architectures
Instructions for analyzing, storing, and delivering streaming data
Tips on aggregating data and working with sets
Information on data warehousing options and techniques
Real-Time Analytics includes in-depth case studies for website analytics, Big Data, visualizing streaming and mobile data, and mining and visualizing operational data flows. The book's "recipe" layout lets readers quickly learn and implement different techniques. All of the code examples presented in the book, along with their related data sets, are available on the companion website.

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

This book introduces you to the Big Data processing techniques addressing but not limited to various BI (business intelligence) requirements, such as reporting, batch analytics, online analytical processing (OLAP), data mining and Warehousing, and predictive analytics. The book has been written on IBM's Platform of Hadoop framework. IBM InfoSphere BigInsight has the highest amount of tutorial matter available free of cost on Internet which makes it easy to acquire proficiency in this technique. This therefore becomes highly vulnerable coaching materials in easy to learn steps. The book optimally provides the courseware as per MCA and M. Tech Level Syllabi of most of the Universities. All components of big Data Platform like Jaql, Hive Pig, Sqoop, Flume, Hadoop Streaming, Oozie: HBase, HDFS, FlumeNG, Whirr, Cloudera, Fuse, Zookeeper and Mahout: Machine learning for Hadoop has been discussed in sufficient Detail with hands on Exercises on each.

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.

Big data has more disruptive potential than any information technology developed in the past 40 years. As author Jeffrey Needham points out in this revealing book, big data can provide unprecedented visibility into the operational efficiency of enterprises and agencies. Disruptive Possibilities provides an historically-informed overview through a wide range of topics, from the evolution of commodity supercomputing and the simplicity of big data technology, to the ways conventional clouds differ from Hadoop analytics clouds. This relentlessly innovative form of computing will soon become standard practice for organizations of any size attempting to derive insight from the tsunami of data engulfing them. Replacing legacy silos—whether they're infrastructure, organizational, or vendor silos—with a platform-centric perspective is just one of the big stories of big data. To reap maximum value from the myriad forms of data, organizations and vendors will have to adopt highly collaborative habits and methodologies.

The data lake is a daring new approach for harnessing the power of big data technology and providing convenient self-service capabilities. But is it right for your company? This book is based on discussions with practitioners and executives from more than a hundred organizations, ranging from data-driven companies such as Google, LinkedIn, and Facebook, to governments and traditional corporate enterprises. You'll learn what a data lake is, why enterprises need one, and how to build one successfully with the best practices in this book. Alex Gorelik, CTO and founder of Waterline Data, explains why old systems and processes can no longer support data needs in the enterprise. Then, in a collection of essays about data lake implementation, you'll examine data lake initiatives, analytic projects, experiences, and best practices from data experts working in various industries. Get a succinct introduction to data warehousing, big data, and data science Learn various paths enterprises take to build a data lake Explore how to build a self-service model and best practices for providing analysts access to the data Use different methods for architecting your data lake Discover ways to implement a data lake from experts in different industries This volume represents the proceedings of the 7th International Conference on Innovation, Communication and Engineering (ICICE 2018), which was held in

P.R. China, November 9-14, 2018. The conference aimed to provide an integrated communication platform for researchers in a wide range of fields including information technology, communication science, applied mathematics, computer science, advanced material science, and engineering. Hopefully, the conference and resulting proceedings will enhance interdisciplinary collaborations between science and engineering technologists in academia and industry within this unique international network.

A practical guide to implementing your enterprise data lake using Lambda Architecture as the base About This Book Build a full-fledged data lake for your organization with popular big data technologies using the Lambda architecture as the base Delve into the big data technologies required to meet modern day business strategies A highly practical guide to implementing enterprise data lakes with lots of examples and real-world use-cases Who This Book Is For Java developers and architects who would like to implement a data lake for their enterprise will find this book useful. If you want to get hands-on experience with the Lambda Architecture and big data technologies by implementing a practical solution using these technologies, this book will also help you. What You Will Learn Build an enterprise-level data lake using the relevant big data technologies Understand the core of the Lambda architecture and how to apply it in an enterprise Learn the technical details around Sqoop and its functionalities Integrate Kafka with Hadoop components to acquire enterprise data Use flume with streaming technologies for stream-based processing Understand stream-based processing with reference to Apache Spark Streaming Incorporate Hadoop components and know the advantages they provide for enterprise data lakes Build fast, streaming, and high-performance applications using ElasticSearch Make your data ingestion process consistent across various data formats with configurability Process your data to derive intelligence using machine learning algorithms In Detail The term "Data Lake" has recently emerged as a prominent term in the big data industry. Data scientists can make use of it in deriving meaningful insights that can be used by businesses to redefine or transform the way they operate. Lambda architecture is also emerging as one of the very eminent patterns in the big data landscape, as it not only helps to derive useful information from historical data but also correlates real-time data to enable business to take critical decisions. This book tries to bring these two important aspects — data lake and lambda architecture—together. This book is divided into three main sections. The first introduces you to the concept of data lakes, the importance of data lakes in enterprises, and getting you up-to-speed with the Lambda architecture. The second section delves into the principal components of building a data lake using the Lambda architecture. It introduces you to popular big data technologies such as Apache Hadoop, Spark, Sqoop, Flume, and ElasticSearch. The third section is a highly practical demonstration of putting it all together, and shows you how an enterprise data lake can be implemented, along with several real-world use-cases. It also shows you how other peripheral

components can be added to the lake to make it more efficient. By the end of this book, you will be able to choose the right big data technologies using the lambda architectural patterns to build your enterprise data lake. Style and approach The book takes a pragmatic approach, showing ways to leverage big data technologies and lambda architecture to build an enterprise-level data lake. Applications of Big Data in Healthcare: Theory and Practice begins with the basics of Big Data analysis and introduces the tools, processes and procedures associated with Big Data analytics. The book unites healthcare with Big Data analysis and uses the advantages of the latter to solve the problems faced by the former. The authors present the challenges faced by the healthcare industry, including capturing, storing, searching, sharing and analyzing data. This book illustrates the challenges in the applications of Big Data and suggests ways to overcome them, with a primary emphasis on data repositories, challenges, and concepts for data scientists, engineers and clinicians. The applications of Big Data have grown tremendously within the past few years and its growth can not only be attributed to its competence to handle large data streams but also to its abilities to find insights from complex, noisy, heterogeneous, longitudinal and voluminous data. The main objectives of Big Data in the healthcare sector is to come up with ways to provide personalized healthcare to patients by taking into account the enormous amounts of already existing data. Provides case studies that illustrate the business processes underlying the use of big data and deep learning health analytics to improve health care delivery Supplies readers with a foundation for further specialized study in clinical analysis and data management Includes links to websites, videos, articles and other online content to expand and support the primary learning objectives for each major section of the book Hadoop in Practice Simon and Schuster

This easy-to-read textbook/reference presents a comprehensive introduction to databases, opening with a concise history of databases and of data as an organisational asset. As relational database management systems are no longer the only database solution, the book takes a wider view of database technology, encompassing big data, NoSQL, object and object-relational and in-memory databases. The text also examines the issues of scalability, availability, performance and security encountered when building and running a database in the real world. Topics and features: presents review and discussion questions at the end of each chapter, in addition to skill-building, hands-on exercises; introduces the fundamental concepts and technologies in database systems, placing these in an historic context; describes the challenges faced by database professionals; reviews the use of a variety of database types in business environments; discusses areas for further research within this fast-moving domain.

This hands-on guide demonstrates how the flexibility of the command line can help you become a more efficient and productive data scientist. You'll learn how to combine small, yet powerful, command-line tools to quickly obtain, scrub,

explore, and model your data. To get you started—whether you're on Windows, OS X, or Linux—author Jeroen Janssens introduces the Data Science Toolbox, an easy-to-install virtual environment packed with over 80 command-line tools. Discover why the command line is an agile, scalable, and extensible technology. Even if you're already comfortable processing data with, say, Python or R, you'll greatly improve your data science workflow by also leveraging the power of the command line. Obtain data from websites, APIs, databases, and spreadsheets Perform scrub operations on plain text, CSV, HTML/XML, and JSON Explore data, compute descriptive statistics, and create visualizations Manage your data science workflow using Drake Create reusable tools from one-liners and existing Python or R code Parallelize and distribute data-intensive pipelines using GNU Parallel Model data with dimensionality reduction, clustering, regression, and classification algorithms

Essentials of Forensic Accounting Essentials of Forensic Accounting is an authoritative resource covering a comprehensive range of forensic accounting topics. As a foundation review, a reference book, or as preparation for the Certification in Financial Forensics (CFF®) Exam, this publication will provide thoughtful and insightful examination of the key themes in this field, including: Professional responsibilities and practice management Fundamental forensic knowledge including laws, courts, and dispute resolution Specialized forensic knowledge such as bankruptcy, insolvency, reorganization, and valuation Through illustrative examples, cases, and explanations, this book makes abstract concepts come to life to help you understand and successfully navigate this complex area.

Utilize R to uncover hidden patterns in your Big Data About This Book Perform computational analyses on Big Data to generate meaningful results Get a practical knowledge of R programming language while working on Big Data platforms like Hadoop, Spark, H2O and SQL/NoSQL databases, Explore fast, streaming, and scalable data analysis with the most cutting-edge technologies in the market Who This Book Is For This book is intended for Data Analysts, Scientists, Data Engineers, Statisticians, Researchers, who want to integrate R with their current or future Big Data workflows. It is assumed that readers have some experience in data analysis and understanding of data management and algorithmic processing of large quantities of data, however they may lack specific skills related to R. What You Will Learn Learn about current state of Big Data processing using R programming language and its powerful statistical capabilities Deploy Big Data analytics platforms with selected Big Data tools supported by R in a cost-effective and time-saving manner Apply the R language to real-world Big Data problems on a multi-node Hadoop cluster, e.g. electricity consumption across various socio-demographic indicators and bike share scheme usage Explore the compatibility of R with Hadoop, Spark, SQL and NoSQL databases, and H2O platform In Detail Big Data analytics is the process of examining large and complex data sets that often exceed the computational capabilities. R is a leading programming language of data science, consisting of powerful functions to tackle all problems related to Big Data processing. The book will begin with a brief introduction to the Big Data world and its current industry standards. With introduction to the R language and presenting its development, structure, applications in real world, and its shortcomings. Book will progress towards revision of major R functions for data management and transformations. Readers will be introduce to Cloud based Big Data solutions (e.g.

Where To Download Practical Hadoop By Example Nyoug

Amazon EC2 instances and Amazon RDS, Microsoft Azure and its HDInsight clusters) and also provide guidance on R connectivity with relational and non-relational databases such as MongoDB and HBase etc. It will further expand to include Big Data tools such as Apache Hadoop ecosystem, HDFS and MapReduce frameworks. Also other R compatible tools such as Apache Spark, its machine learning library Spark MLlib, as well as H2O. Style and approach This book will serve as a practical guide to tackling Big Data problems using R programming language and its statistical environment. Each section of the book will present you with concise and easy-to-follow steps on how to process, transform and analyse large data sets.

Leverage the capabilities of SAS to process and analyze Big Data About This Book Combine SAS with platforms such as Hadoop, SAP HANA, and Cloud Foundry-based platforms for efficient Big Data analytics Learn how to use the web browser-based SAS Studio and iPython Jupyter Notebook interfaces with SAS Practical, real-world examples on predictive modeling, forecasting, optimizing and reporting your Big Data analysis with SAS Who This Book Is For SAS professionals and data analysts who wish to perform analytics on Big Data using SAS to gain actionable insights will find this book to be very useful. If you are a data science professional looking to perform large-scale analytics with SAS, this book will also help you. A basic understanding of SAS will be helpful, but is not mandatory. What You Will Learn Configure a free version of SAS in order to do hands-on exercises dealing with data management, analysis, and reporting. Understand the basic concepts of the SAS language which consists of the data step (for data preparation) and procedures (or PROCs) for analysis. Make use of the web browser based SAS Studio and iPython Jupyter Notebook interfaces for coding in the SAS, DS2, and FedSQL programming languages. Understand how the DS2 programming language plays an important role in Big Data preparation and analysis using SAS Integrate and work efficiently with Big Data platforms like Hadoop, SAP HANA, and cloud foundry based systems. In Detail SAS has been recognized by Money Magazine and Payscale as one of the top business skills to learn in order to advance one's career. Through innovative data management, analytics, and business intelligence software and services, SAS helps customers solve their business problems by allowing them to make better decisions faster. This book introduces the reader to the SAS and how they can use SAS to perform efficient analysis on any size data, including Big Data. The reader will learn how to prepare data for analysis, perform predictive, forecasting, and optimization analysis and then deploy or report on the results of these analyses. While performing the coding examples within this book the reader will learn how to use the web browser based SAS Studio and iPython Jupyter Notebook interfaces for working with SAS. Finally, the reader will learn how SAS's architecture is engineered and designed to scale up and/or out and be combined with the open source offerings such as Hadoop, Python, and R. By the end of this book, you will be able to clearly understand how you can efficiently analyze Big Data using SAS. Style and approach The book starts off by introducing the reader to SAS and the SAS programming language which provides data management, analytical, and reporting capabilities. Most chapters include hands on examples which highlights how SAS provides The Power to Know®. The reader will learn that if they are looking to perform large-scale data analysis that SAS provides an open platform engineered and designed to scale both up and out which allows the power of SAS to combine with open source offerings such as Hadoop, Python, and R.

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

Where To Download Practical Hadoop By Example Nyoug

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.

Filled with practical, step-by-step instructions and clear explanations for the most important and useful tasks. This is a Packt Instant How-to guide, which provides concise and clear recipes for getting started with Hadoop. This book is for big data enthusiasts and would-be Hadoop programmers. It is also meant for Java programmers who either have not worked with Hadoop at all, or who know Hadoop and MapReduce but are not sure how to deepen their understanding.

This book describes current problems in data science and Big Data. Key topics are data classification, Graph Cut, the Laplacian Matrix, Google Page Rank, efficient algorithms, hardness of problems, different types of big data, geometric data structures, topological data processing, and various learning methods. For unsolved problems such as incomplete data relation and reconstruction, the book includes possible solutions and both statistical and computational methods for data analysis. Initial chapters focus on exploring the properties of incomplete data sets and partial-connectedness among data points or data sets. Discussions also cover the completion problem of Netflix matrix; machine learning method on massive data sets; image segmentation and video search. This book introduces software tools for data science and Big Data such MapReduce, Hadoop, and Spark. This book contains three parts. The first part explores the fundamental tools of data science. It includes basic graph theoretical methods, statistical and AI methods for massive data sets. In second part, chapters focus on the procedural treatment of data science problems including machine learning methods, mathematical image and video processing, topological data analysis, and statistical methods. The final section provides case studies on special topics in variational learning, manifold learning, business and financial data recovery, geometric search, and computing models. Mathematical Problems in Data Science is a valuable resource for researchers and professionals working in data science, information systems and networks. Advanced-level students studying computer science, electrical engineering and mathematics will also find the content helpful.

[Copyright: 98bc745dde909269acb07416978d352a](#)