Big Data Principles And Best Practices Of Scalable Realtime Data Systems

This is likewise one of the factors by obtaining the soft documents of this **Big Data Principles And Best Practices Of Scalable Realtime Data Systems** by online. You might not require more period to spend to go to the books launch as with ease as search for them. In some cases, you likewise get not discover the proclamation Big Data Principles And Best Practices Of Scalable Realtime Data Systems that you are looking for. It will very squander the time.

However below, similar to you visit this web page, it will be so very simple to get as well as download guide Big Data Principles And Best Practices Of Scalable Realtime Data Systems

It will not take many epoch as we accustom before. You can reach it while doing something else at home and even in your workplace. appropriately easy! So, are you question? Just exercise just what we manage to pay for under as with ease as review **Big Data Principles And Best Practices Of Scalable Realtime Data Systems** what you considering to read!

Designing Data-Intensive Applications - Martin Kleppmann 2017-03-16
Data is at the center of many challenges in system design today. Difficult issues need to be figured out, such as scalability, consistency, reliability, efficiency, and maintainability. In addition, we have an overwhelming variety of tools, including relational databases, NoSQL datastores, stream or batch processors, and message brokers. What are the right choices for your application? How do you make sense of all these buzzwords? In this practical and comprehensive guide, author Martin Kleppmann helps you navigate this diverse landscape by examining the pros and cons of various technologies for processing and storing data. Software keeps changing, but the fundamental principles remain the same. With this book, software engineers and architects will learn how to apply those ideas in practice, and how to make full use of data in modern applications. Peer under the hood of the systems you already use, and learn how to use and operate them more effectively Make informed decisions by identifying the strengths and weaknesses of different tools Navigate the trade-offs around consistency, scalability, fault tolerance, and complexity Understand the distributed systems research upon which modern databases are built Peek behind the scenes of major online services, and learn from their architectures

Real-World Machine Learning - Henrik Brink 2016-09-15
Summary Real-World Machine Learning is a practical guide designed to teach working developers the art of ML project execution. Without overdosing you on academic theory and complex mathematics, it introduces the day-to-day practice of machine learning, preparing you to successfully build and deploy powerful ML systems. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Machine learning systems help you find valuable insights and patterns in data, which you'd never recognize with traditional methods. In the real world, ML techniques give you a way to identify trends, forecast behavior, and make fact-based recommendations. It's a hot and growing field, and up-to-speed ML developers are in demand. About the Book Real-World Machine Learning will teach you the concepts and techniques you need to be a successful machine learning practitioner without overdosing you on abstract theory and complex mathematics. By working through immediately relevant examples in Python, you'll build skills in data acquisition and modeling, classification, and regression. You'll also explore the most important tasks like model validation, optimization, scalability, and real-time streaming. When you're done, you'll be ready to successfully build, deploy, and maintain your own powerful ML systems. What's Inside Predicting future behavior Performance evaluation and optimization Analyzing sentiment and making recommendations About the Reader No prior machine learning experience assumed. Readers should know Python. About the Authors Henrik Brink, Joseph Richards and Mark Fetherolf are experienced data scientists engaged in the daily practice of machine learning. Table of Contents PART 1: THE MACHINE-LEARNING WORKFLOW What is machine learning? Real-world data Modeling and prediction Model evaluation and optimization Basic feature engineering PART 2: PRACTICAL APPLICATION Example: NYC taxi data Advanced feature engineering Advanced NLP example: movie review sentiment Scaling machine-learning workflows Example: digital display advertising

Privacy, Big Data, and the Public Good - Julia Lane 2014-06-09
Massive amounts of data on human beings can now be analyzed. Pragmatic purposes abound, including selling goods and services, winning political campaigns, and identifying possible terrorists. Yet 'big data' can also be harnessed to serve the public good: scientists can use big data to do research that improves the lives of human beings, improves government services, and reduces taxpayer costs. In order to achieve this goal, researchers must have access to this data - raising important privacy questions. What are the ethical and legal requirements? What are the rules of engagement? What are the best ways to provide access while also protecting confidentiality? Are there reasonable mechanisms to compensate citizens for privacy loss? The goal of this book is to answer some of these questions. The book's authors paint an intellectual landscape that includes legal, economic, and statistical frameworks. The authors also identify new practical approaches that simultaneously maximize the utility of data access while minimizing information risk.

Big Data Fundamentals - Thomas Erl 2015-12-29
“This text should be required reading for everyone in contemporary business.” --Peter Woodhull, CEO, Modus21 “The one book that clearly describes and links Big Data concepts to business utility.” --Dr. Christopher Starr, PhD “Simply, this is the best Big Data book on the market!” --Sam Rostam, Cascadian IT Group “...one of the most contemporary approaches I’ve seen to Big Data fundamentals...” --Joshua M. Davis, PhD The Definitive Plain-English Guide to Big Data for Business and Technology Professionals Big Data Fundamentals provides a pragmatic, no-nonsense introduction to Big Data. Best-selling IT author Thomas Erl and his team clearly explain key Big Data concepts, theory and terminology, as well as fundamental technologies and techniques. All coverage is supported with case study examples and numerous simple diagrams. The authors begin by explaining how Big Data can propel an organization forward by solving a spectrum of previously intractable business problems. Next, they demystify key analysis techniques and technologies and show how a Big Data solution environment can be built and integrated to offer competitive advantages. Discovering Big Data’s fundamental concepts and what makes it different from previous forms of data analysis and data science Understanding the business motivations and drivers behind Big Data adoption, from operational improvements through innovation Planning strategic, business-driven Big Data initiatives Addressing considerations such as data management, governance, and security Recognizing the 5 “V” characteristics of datasets in Big Data environments: volume, velocity, variety, veracity, and value Clarifying Big Data’s relationships with OLTP, OLAP, ETL, data warehouses, and data marts Working with Big Data in structured, unstructured, semi-structured, and metadata formats Increasing value by integrating Big Data resources with corporate performance monitoring Understanding how Big Data leverages distributed and parallel processing Using NoSQL and other technologies to meet Big Data’s distinct data processing requirements Leveraging statistical approaches of quantitative and qualitative analysis Applying computational analysis methods, including machine learning

Big Data - Nathan Marz 2015
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

Mastering Spark with R - Javier Luraschi 2019-10-07

If you’re like most R users, you have deep knowledge and love for statistics. But as your organization continues to collect huge amounts of data, adding tools such as Apache Spark makes a lot of sense. With this practical book, data scientists and professionals working with large-scale data applications will learn how to use Spark from R to tackle big data and big compute problems. Authors Javier Luraschi, Kevin Kuo, and Edgar Ruiz show you how to use R with Spark to solve different data analysis problems. This book covers relevant data science topics, cluster computing, and issues that should interest even the most advanced users. Analyze, explore, transform, and visualize data in Apache Spark with R Create statistical models to extract information and predict outcomes; automate the process in production-ready workflows Perform analysis and modeling across many machines using distributed computing techniques Use large-scale data from multiple sources and different formats with ease from within Spark Learn about alternative modeling frameworks for graph processing, geospatial analysis, and genomics at scale Dive into advanced topics including custom transformations, real-time data processing, and creating custom Spark extensions

Big Data Management - Peter Ghavami 2020-11-09

Data analytics is core to business and decision making. The rapid increase in data volume, velocity and variety offers both opportunities and challenges. While open source solutions to store big data, like Hadoop, offer platforms for exploring value and insight from big data, they were not originally developed with data security and governance in mind. Big Data Management discusses numerous policies, strategies and recipes for managing big data. It addresses data security, privacy, controls and life cycle management offering modern principles and open source architectures for successful governance of big data. The author has collected best practices from the world’s leading organizations that have successfully implemented big data platforms. The topics discussed cover the entire data management life cycle, data quality, data stewardship, regulatory considerations, data council, architectural and operational models are presented for successful management of big data. The book is a must-read for data scientists, data engineers and corporate leaders who are implementing big data platforms in their organizations.

97 Things Every Data Engineer Should Know - Tobias Macey 2021-06-11

Take advantage of today's sky-high demand for data engineers. With this in-depth book, current and aspiring engineers will learn powerful real-world best practices for managing data big and small. Contributors from notable companies including Twitter, Google, Stitch Fix, Microsoft, Capital One, and LinkedIn share their experiences and lessons learned for overcoming a variety of specific and often nagging challenges. Edited by Tobias Macey, host of the popular Data Engineering Podcast, this book presents 97 concise and useful tips for cleaning, prepping, wrangling, storing, processing, and ingesting data. Data engineers, data architects, data team managers, data scientists, machine learning engineers, and software engineers will greatly benefit from the wisdom and experience of their peers. Topics include: The Importance of Data Lineage - Julien Le Dem Data Security for Data Engineers - Katharine Jarmul The Two Types of Data Engineering and Data Engineers - Jesse Anderson Six Dimensions for Picking an Analytical Data Warehouse - Gleb Mezhanskiy The End of ETL as We Know It - Paul Singman Building a Career as a Data Engineer - Vijay Kiran Modern Metadata for the Modern Data Stack - Prukalpa Sankar Your Data Tests Failed! Now What? - Sam Bail

Big Data Imperatives - Soumendra Mohanty 2013-08-23

Big Data Imperatives, focuses on resolving the key questions on everyone’s mind: Which data matters? Do you have enough data volume to justify the usage? How you want to process this amount of data? How long do you really need to keep it active for your analysis, marketing, and BI applications? Big data is emerging from the realm of one-off projects to mainstream business adoption; however, the real value of big data is not in the overwhelming size of it, but more in its effective use. This book addresses the following big data characteristics: Very large, distributed aggregations of loosely structured data – often incomplete and inaccessible Petabytes/Exabytes of data Millions/billions of people providing/contributing to the context behind the data Flat schema's with few complex interrelationships Involves time-stamped events Made up of incomplete data Includes connections between data elements that must be probabilistically inferred Big Data Imperatives explains 'what big data can do'. It can batch process millions and billions of records both unstructured and structured much faster and cheaper. Big data analytics provide a platform to merge all analysis which enables data analysis to be more accurate, well-rounded, reliable and focused on a specific business capability. Big Data Imperatives describes the complementary nature of traditional data warehouses and big-data analytics platforms and how they feed each other. This book aims to bring the big data and analytics realms together with a greater focus on architectures that leverage the scale and power of big data and the ability to integrate and apply analytics principles to data which earlier was not accessible. This book can also be used as a handbook for practitioners; helping them on methodology,technical architecture, analytics techniques and best practices. At the same time, this book intends to hold the interest of those new to big data and analytics by giving them a deep insight into the realm of big data.

Developing Data Migrations and Integrations with Salesforce - David Masri 2018-12-18

Migrate your data to Salesforce and build low-maintenance and high-performing data integrations to get the most out of Salesforce and make it a "go-to" place for all your organization's customer information. When companies choose to roll out Salesforce, users expect it to be the place to find any and all Information related to a customer—the coveted Client 360° view. On the day you go live, users expect to see all their accounts, contacts, and historical data in the system. They also expect that data entered in other systems will be exposed in Salesforce automatically and in a timely manner. This book shows you how to migrate all your legacy data to Salesforce and then design integrations to your organization's mission-critical systems. As the Salesforce platform grows more powerful, it also grows in complexity. Whether you are migrating data to Salesforce, or integrating with Salesforce, it is important to understand how these complexities need to be reflected in your design. Developing Data Migrations and Integrations with Salesforce covers everything you need to know to migrate your data to Salesforce the right way, and how to design low-maintenance, high-performing data integrations with Salesforce. This book is written by a practicing Salesforce integration architect with dozens of Salesforce projects under his belt. The patterns and practices covered in this book are the results of the lessons learned during those projects. What You’ll Learn Know how Salesforce’s data engine is architected and why Use the Salesforce Data APIs to load and extract data Plan and execute your data migration to Salesforce Design low-maintenance, high-performing data integrations with Salesforce Understand common data integration patterns and the pros and cons of each Know real-time integration options for Salesforce Be aware of common pitfalls Build reusable transformation code covering commonly needed Salesforce transformation patterns Who This Book Is For Those tasked with migrating data to Salesforce or building ongoing data

integrations with Salesforce, regardless of the ETL tool or middleware chosen; project sponsors or managers nervous about data tracks putting their projects at risk; aspiring Salesforce integration and/or migration specialists; Salesforce developers or architects looking to expand their skills and take on new challenges

Data Engineering - Brian Shive 2013-10-01

If you found a rusty old lamp on the beach, and upon touching it a genie appeared and granted you three wishes, what would you wish for? If you were wishing for a successful application development effort, most likely you would wish for accurate and robust data models, comprehensive data flow diagrams, and an acute understanding of human behavior. The wish for well-designed conceptual and logical data models means the requirements are well-understood and that the design has been built with flexibility and extensibility leading to high application agility and low maintenance costs. The wish for detailed data flow diagrams means a concrete understanding of the business' value chain exists and is documented. The wish to understand how we think means excellent team dynamics while analyzing, designing, and building the application. Why search the beaches for genie lamps when instead you can read this book? Learn the skills required for modeling, value chain analysis, and team dynamics by following the journey the author and son go through in establishing a profitable summer lemonade business. This business grew from season to season proportionately with his adoption of important engineering principles. All of the concepts and principles are explained in a novel format, so you will learn the important messages while enjoying the story that unfolds within these pages. The story is about an old man who has spent his life designing data models and databases and his newly adopted son. Father and son have a 54 year age difference that produces a large generation gap. The father attempts to narrow the generation gap by having his nine-year-old son earn his entertainment money. The son must run a summer business that turns a lemon grove into profits so he can buy new computers and games. As the son struggles for profits, it becomes increasingly clear that dad's career in information technology can provide critical leverage in achieving success in business. The failures and successes of the son's business over the summers are a microcosm of the ups and downs of many enterprises as they struggle to manage information technology.

The Enterprise Big Data Lake - Alex Gorelik 2019-02-21

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

Applied Predictive Analytics - Dean Abbott 2014-04-14

Learn the art and science of predictive analytics — techniques that get results Predictive analytics is what translates big data into meaningful, usable business information. Written by a leading expert in the field, this guide examines the science of the underlying algorithms as well as the principles and best practices that govern the art of predictive analytics. It clearly explains the theory behind predictive analytics, teaches the methods, principles, and techniques for conducting predictive analytics projects, and offers tips and tricks that are essential for successful predictive modeling. Hands-on examples and case studies are included. The ability to successfully apply predictive analytics enables businesses to effectively interpret big data; essential for competition today This guide teaches not only the principles of predictive analytics, but also how to apply them to achieve real, pragmatic solutions Explains methods, principles, and techniques for conducting predictive analytics projects from start to finish Illustrates each technique with hands-on examples and includes as series of in-depth case studies that apply predictive analytics to common business scenarios A companion website provides all the data sets used to generate the examples as well as a free trial version of software Applied Predictive Analytics arms data and business analysts and business managers with the tools they need to interpret and capitalize on big data.

Big Data For Dummies - Judith S. Hurwitz 2013-04-02

Find the right big data solution for your business ororganization Big data management is one of the major challenges facingbusiness, industry, and not-for-profit organizations. Data setssuch as customer transactions for a mega-retailer, weather patternsmonitored by meteorologists, or social network activity can quicklyoutpace the capacity of traditional data management tools. If youneed to develop or manage big data solutions, you'll appreciate howthese four experts define, explain, and guide you through this newand often confusing concept. You'll learn what it is, why itmatters, and how to choose and implement solutions that work. Effectively managing big data is an issue of growing importanceto businesses, not-for-profit organizations, government, and ITprofessionals Authors are experts in information management, big data, and avariety of solutions Explains big data in detail and discusses how to select andimplement a solution, security concerns to consider, data storageand presentation issues, analytics, and much more Provides essential information in a no-nonsense,easy-to-understand style that is empowering Big Data For Dummies cuts through the confusion and helpsyou take charge of big data solutions for your organization.

Data Architecture: A Primer for the Data Scientist - W.H. Inmon 2014-11-26

Today, the world is trying to create and educate data scientists because of the phenomenon of Big Data. And everyone is looking deeply into this technology. But no one is looking at the larger architectural picture of how Big Data needs to fit within the existing systems (data warehousing systems). Taking a look at the larger picture into which Big Data fits gives the data scientist the necessary context for how pieces of the puzzle should fit together. Most references on Big Data look at only one tiny part of a much larger whole. Until data gathered can be put into an existing framework or architecture it can't be used to its full potential. Data Architecture a Primer for the Data Scientist addresses the larger architectural picture of how Big Data fits with the existing information infrastructure, an essential topic for the data scientist. Drawing upon years of practical experience and using numerous examples and an easy to understand framework. W.H. Inmon, and Daniel Linstedt define the importance of data architecture and how it can be used effectively to harness big data within existing systems. You'll be able to: Turn textual information into a form that can be analyzed by standard tools. Make the connection between analytics and Big Data Understand how Big Data fits within an existing systems environment Conduct analytics on repetitive and non-repetitive data Discusses the value in Big Data that is often overlooked, non-repetitive data, and why there is significant business value in using it Shows how to turn textual information into a form that can be analyzed by standard tools Explains how Big Data fits within an existing systems environment Presents new opportunities that are afforded by the advent of Big Data Demystifies the murky waters of repetitive and non-repetitive data in Big Data

Data Privacy - Nataraj Venkataramanan 2016-10-03

The book covers data privacy in depth with respect to data mining, test data management, synthetic data generation etc. It formalizes principles of data privacy that are essential for good anonymization design based on the data format and discipline. The principles outline best practices and reflect on the conflicting relationship between privacy and utility. From a practice standpoint, it provides practitioners and researchers with a definitive guide to approach anonymization of various data formats, including multidimensional, longitudinal, time-series, transaction, and graph data. In addition to helping CIOs protect confidential data, it also offers a guideline as to how this can be implemented for a wide range of data at the enterprise level.

Big Data - Rajkumar Buyya 2016-06-07

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

next generation Big Data platforms Explores the challenges in Big Data computing and ways to overcome them Contains expert contributors from both academia and industry

Data Engineering with Python - Paul Crickard 2020-10-23

Build, monitor, and manage real-time data pipelines to create data engineering infrastructure efficiently using open-source Apache projects Key FeaturesBecome well-versed in data architectures, data preparation, and data optimization skills with the help of practical examplesDesign data models and learn how to extract, transform, and load (ETL) data using PythonSchedule, automate, and monitor complex data pipelines in productionBook Description Data engineering provides the foundation for data science and analytics, and forms an important part of all businesses. This book will help you to explore various tools and methods that are used for understanding the data engineering process using Python. The book will show you how to tackle challenges commonly faced in different aspects of data engineering. You'll start with an introduction to the basics of data engineering, along with the technologies and frameworks required to build data pipelines to work with large datasets. You'll learn how to transform and clean data and perform analytics to get the most out of your data. As you advance, you'll discover how to work with big data of varying complexity and production databases, and build data pipelines. Using real-world examples, you'll build architectures on which you'll learn how to deploy data pipelines. By the end of this Python book, you'll have gained a clear understanding of data modeling techniques, and will be able to confidently build data engineering pipelines for tracking data, running quality checks, and making necessary changes in production. What you will learnUnderstand how data engineering supports data science workflowsDiscover how to extract data from files and databases and then clean, transform, and enrich itConfigure processors for handling different file formats as well as both relational and NoSQL databasesFind out how to implement a data pipeline and dashboard to visualize resultsUse staging and validation to check data before landing in the warehouseBuild real-time pipelines with staging areas that perform validation and handle failuresGet to grips with deploying pipelines in the production environmentWho this book is for This book is for data analysts, ETL developers, and anyone looking to get started with or transition to the field of data engineering or refresh their knowledge of data engineering using Python. This book will also be useful for students planning to build a career in data engineering or IT professionals preparing for a transition. No previous knowledge of data engineering is required.

Big Data, Analytics, and the Future of Marketing & Sales - McKinsey Chief McKinsey Chief Marketing & Sales Officer Forum 2014-08-16

Big Data is the biggest game-changing opportunity for marketing and sales since the Internet went mainstream almost 20 years ago. The data big bang has unleashed torrents of terabytes about everything from customer behaviors to weather patterns to demographic consumer shifts in emerging markets. This collection of articles, videos, interviews, and slideshares highlights the most important lessons for companies looking to turn data into above-market growth: Using analytics to identify valuable business opportunities from the data to drive decisions and improve marketing return on investment (MROI) Turning those insights into well-designed products and offers that delight customers Delivering those products and offers effectively to the marketplace.The goldmine of data represents a pivot-point moment for marketing and sales leaders. Companies that inject big data and analytics into their operations show productivity rates and profitability that are 5 percent to 6 percent higher than those of their peers. That's an advantage no company can afford to ignore.

Big Data and Social Science - Ian Foster 2020-11-17

Big Data and Social Science: Data Science Methods and Tools for Research and Practice, Second Edition shows how to apply data science to real-world problems, covering all stages of a data-intensive social science or policy project. Prominent leaders in the social sciences, statistics, and computer science as well as the field of data science provide a unique perspective on how to apply modern social science research principles and current analytical and computational tools. The text teaches you how to identify and collect appropriate data, apply data science methods and tools to the data, and recognize and respond to data errors, biases, and limitations. Features: Takes an accessible, hands-on approach to handling new types of data in the social sciences Presents the key data science tools in a non-intimidating way to both social and data scientists while keeping the focus on research questions and purposes Illustrates social science and data science principles through real-world problems Links computer science concepts to practical social science research Promotes good scientific practice Provides freely available workbooks with data, code, and practical programming exercises, through Binder and GitHub New to the Second Edition: Increased use of examples from different areas of social sciences New chapter on dealing with Bias and Fairness in Machine Learning models Expanded chapters focusing on Machine Learning and Text Analysis Revamped hands-on Jupyter notebooks to reinforce concepts covered in each chapter This classroom-tested book fills a major gap in graduate- and professional-level data science and social science education. It can be used to train a new generation of social data scientists to tackle real-world problems and improve the skills and competencies of applied social scientists and public policy practitioners. It empowers you to use the massive and rapidly growing amounts of available data to interpret economic and social activities in a scientific and rigorous manner.

Applied Data Science - Martin Braschler 2019-06-13

This book has two main goals: to define data science through the work of data scientists and their results, namely data products, while simultaneously providing the reader with relevant lessons learned from applied data science projects at the intersection of academia and industry. As such, it is not a replacement for a classical textbook (i.e., it does not elaborate on fundamentals of methods and principles described elsewhere), but systematically highlights the connection between theory, on the one hand, and its application in specific use cases, on the other. With these goals in mind, the book is divided into three parts: Part I pays tribute to the interdisciplinary nature of data science and provides a common understanding of data science terminology for readers with different backgrounds. These six chapters are geared towards drawing a consistent picture of data science and were predominantly written by the editors themselves. Part II then broadens the spectrum by presenting views and insights from diverse authors – some from academia and some from industry, ranging from financial to health and from manufacturing to e-commerce. Each of these chapters describes a fundamental principle, method or tool in data science by analyzing specific use cases and drawing concrete conclusions from them. The case studies presented, and the methods and tools applied, represent the nuts and bolts of data science. Finally, Part III was again written from the perspective of the editors and summarizes the lessons learned that have been distilled from the case studies in Part II. The section can be viewed as a meta-study on data science across a broad range of domains, viewpoints and fields. Moreover, it provides answers to the question of what the mission-critical factors for success in different data science undertakings are. The book targets professionals as well as students of data science: first, practicing data scientists in industry and academia who want to broaden their scope and expand their knowledge by drawing on the authors' combined experience. Second, decision makers in businesses who face the challenge of creating or implementing a data-driven strategy and who want to learn from success stories spanning a range of industries. Third, students of data science who want to understand both the theoretical and practical aspects of data science, vetted by real-world case studies at the intersection of academia and industry.

Data Science and Big Data Analytics - EMC Education Services 2015-01-05

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

Knowledge of the Law in the Big Data Age - G. Peruginelli 2019-07-23

The changes brought about by digital technology and the consequent explosion of information known as Big Data have brought opportunities and challenges in all areas of society, and the law is no exception. This book, Knowledge of the Law in the Big Data Age contains a selection of the papers presented at the conference 'Law via the Internet 2018', held

in Florence, Italy, on 11-12 October 2018. This annual conference of the 'Free Access to Law Movement' (http://www.fatlm.org) hosted more than 60 international speakers from universities, government and research bodies as well as EU institutions. Topics covered range from free access to law and Big Data and data analytics in the legal domain, to policy issues concerning access, publishing and the dissemination of legal information, tools to support democratic participation and opportunities for digital democracy. The book is divided into 3 sections: Part I provides an introductory background, covering aspects such as the evolution of legal science and models for representing the law; Part II addresses the present and future of access to law and to various legal information sources; and Part III covers updates in projects, initiatives, and concrete achievements in the field. The book provides an overview of the practical implementation of legal information systems and the tools to manage this special kind of information, as well as some of the critical issues which must be faced, and will be of interest to all those working at the intersection of law and technology.

Information Governance Principles and Practices for a Big Data Landscape - Chuck Ballard 2014-03-31

This IBM® Redbooks® publication describes how the IBM Big Data Platform provides the integrated capabilities that are required for the adoption of Information Governance in the big data landscape. As organizations embark on new use cases, such as Big Data Exploration, an enhanced 360 view of customers, or Data Warehouse modernization, and absorb ever growing volumes and variety of data with accelerating velocity, the principles and practices of Information Governance become ever more critical to ensure trust in data and help organizations overcome the inherent risks and achieve the wanted value. The introduction of big data changes the information landscape. Data arrives faster than humans can react to it, and issues can quickly escalate into significant events. The variety of data now poses new privacy and security risks. The high volume of information in all places makes it harder to find where these issues, risks, and even useful information to drive new value and revenue are. Information Governance provides an organization with a framework that can align their wanted outcomes with their strategic management principles, the people who can implement those principles, and the architecture and platform that are needed to support the big data use cases. The IBM Big Data Platform, coupled with a framework for Information Governance, provides an approach to build, manage, and gain significant value from the big data landscape.

The Big Book of Dashboards - Steve Wexler 2017-04-24

The definitive reference book with real-world solutions you won't find anywhere else The Big Book of Dashboards presents a comprehensive reference for those tasked with building or overseeing the development of business dashboards. Comprising dozens of examples that address different industries and departments (healthcare, transportation, finance, human resources, marketing, customer service, sports, etc.) and different platforms (print, desktop, tablet, smartphone, and conference room display) The Big Book of Dashboards is the only book that matches great dashboards with real-world business scenarios. By organizing the book based on these scenarios and offering practical and effective visualization examples, The Big Book of Dashboards will be the trusted resource that you open when you need to build an effective business dashboard. In addition to the scenarios there's an entire section of the book that is devoted to addressing many practical and psychological factors you will encounter in your work. It's great to have theory and evidenced-based research at your disposal, but what will you do when somebody asks you to make your dashboard 'cooler' by adding packed bubbles and donut charts? The expert authors have a combined 30-plus years of hands-on experience helping people in hundreds of organizations build effective visualizations. They have fought many 'best practices' battles and having endured bring an uncommon empathy to help you, the reader of this book, survive and thrive in the data visualization world. A well-designed dashboard can point out risks, opportunities, and more; but common challenges and misconceptions can make your dashboard useless at best, and misleading at worst. The Big Book of Dashboards gives you the tools, guidance, and models you need to produce great dashboards that inform, enlighten, and engage.

Principles of Database Management - Wilfried Lemahieu 2018-07-12

Introductory, theory-practice balanced text teaching the fundamentals of databases to advanced undergraduates or graduate students in information systems or computer science.

Big Data Governance - Peter Ghavami, Ph.d. 2015-11-26

Data is the new Gold and Analytics is the machinery to mine, mold and mint it. Data analytics has become core to business and decision making. The rapid increase in data volume, velocity and variety, known as big data, offers both opportunities and challenges. While open source solutions to store big data, like Hadoop and NoSQL offer platforms for exploring value and insight from big data, they were not originally developed with data security and governance in mind. Organizations that are launching big data initiatives face significant challenges for managing this data effectively. In this book, the author has collected best practices from the world's leading organizations who have successfully implemented big data platforms. He offers the latest techniques and methods for managing big data effectively. The book offers numerous policies, strategies and recipes for managing big data. It addresses many issues that are prevalent with data security, privacy, controls and life cycle management offering modern principles and open source architectures for successful governance of big data. Topics that cover the entire data management life cycle, data quality, data stewardship, regulatory considerations, data council, architectural and operational models are presented for successful management of big data. The book is a must-read for data scientists, data engineers and information technology leaders who are implementing big data platforms in their organizations.

Managerial Analytics - Michael Watson 2013

The field of analytics is rapidly evolving, making it difficult for professionals and students to keep up the most current and effective applications. Managerial Analytics will help readers sort through all these new options and identify the appropriate solution. In this reference, authors Watson, Nelson and Cacioppi accurately define and identify the components of analytics and big data, giving readers the knowledge needed to effectively assess new aspects and applications. Building on this foundation, they review tools and solutions, identify the offerings best aligned to one's requirements, and show how to tailor analytics applications to an organization's specific needs. Drawing on extensive experience implementing, planning, and researching advanced analytics for business, the authors clearly explain all this, and more: What analytics is and isn't: great examples of successful usage – and other examples where the term is being degraded into meaninglessness The difference between using analytics and "competing on analytics" How to get started with big data, by analyzing the most relevant data Components of analytics systems, from databases and Excel to BI systems and beyond Anticipating and overcoming "confirmation bias" and other pitfalls Understanding predictive analytics and getting the high-quality random samples necessary Applying game theory, Efficient Frontier, benchmarking, and revenue management models Implementing optimization at the small and large scale, and using it to make "automatic decisions"

The Politics and Policies of Big Data - Ann Rudinow Sætnan 2018-05-08

Big Data, gathered together and re-analysed, can be used to form endless variations of our persons - so-called 'data doubles'. Whilst never a precise portrayal of who we are, they unarguably contain glimpses of details about us that, when deployed into various routines (such as management, policing and advertising) can affect us in many ways. How are we to deal with Big Data? When is it beneficial to us? When is it harmful? How might we regulate it? Offering careful and critical analyses, this timely volume aims to broaden well-informed, unprejudiced discourse, focusing on: the tenets of Big Data, the politics of governance and regulation; and Big Data practices, performance and resistance. An interdisciplinary volume, The Politics of Big Data will appeal to undergraduate and postgraduate students, as well as postdoctoral and senior researchers interested in fields such as Technology, Politics and Surveillance.

New Technologies for Human Rights Law and Practice - Molly K. Land 2018-04-19

Provides a roadmap for understanding the relationship between technology and human rights law and practice. This title is also available as Open Access.

Principles of Data Wrangling - Tye Rattenbury 2017-06-29

A key task that any aspiring data-driven organization needs to learn is data wrangling, the process of converting raw data into something truly useful. This practical guide provides business analysts with an overview of various data wrangling techniques and tools, and puts the practice of data wrangling into context by asking, "What are you trying to do and why?" Wrangling data consumes roughly 50-80% of an analyst's time before any kind of analysis is possible. Written by key executives at Trifacta, this book walks you through the wrangling process by exploring several factors—time, granularity, scope, and structure—that you need to

consider as you begin to work with data. You'll learn a shared language and a comprehensive understanding of data wrangling, with an emphasis on recent agile analytic processes used by many of today's data-driven organizations. Appreciate the importance—and the satisfaction—of wrangling data the right way. Understand what kind of data is available Choose which data to use and at what level of detail Meaningfully combine multiple sources of data Decide how to distill the results to a size and shape that can drive downstream analysis

Big Data - Viktor Mayer-Schönberger 2013

This revelatory exploration of big data, which refers to our newfound ability to crunch vast amounts of information, analyze it instantly and draw profound and surprising conclusions from it, discusses how it will change our lives and what we can do to protect ourselves from its hazards. 75,000 first printing.

Knowledge Graphs and Big Data Processing - Valentina Janev 2020-01-01

This open access book is part of the LAMBDA Project (Learning, Applying, Multiplying Big Data Analytics), funded by the European Union, GA No. 809965. Data Analytics involves applying algorithmic processes to derive insights. Nowadays it is used in many industries to allow organizations and companies to make better decisions as well as to verify or disprove existing theories or models. The term data analytics is often used interchangeably with intelligence, statistics, reasoning, data mining, knowledge discovery, and others. The goal of this book is to introduce some of the definitions, methods, tools, frameworks, and solutions for big data processing, starting from the process of information extraction and knowledge representation, via knowledge processing and analytics to visualization, sense-making, and practical applications. Each chapter in this book addresses some pertinent aspect of the data processing chain, with a specific focus on understanding Enterprise Knowledge Graphs, Semantic Big Data Architectures, and Smart Data Analytics solutions. This book is addressed to graduate students from technical disciplines, to professional audiences following continuous education short courses, and to researchers from diverse areas following self-study courses. Basic skills in computer science, mathematics, and statistics are required.

Machine Learning Models and Algorithms for Big Data Classification - Shan Suthaharan 2015-10-20

This book presents machine learning models and algorithms to address big data classification problems. Existing machine learning techniques like the decision tree (a hierarchical approach), random forest (an ensemble hierarchical approach), and deep learning (a layered approach) are highly suitable for the system that can handle such problems. This book helps readers, especially students and newcomers to the field of big data and machine learning, to gain a quick understanding of the techniques and technologies; therefore, the theory, examples, and programs (Matlab and R) presented in this book have been simplified, hardcoded, repeated, or spaced for improvements. They provide vehicles to test and understand the complicated concepts of various topics in the field. It is expected that the readers adopt these programs to experiment with the examples, and then modify or write their own programs toward advancing their knowledge for solving more complex and challenging problems. The presentation format of this book focuses on simplicity, readability, and dependability so that both undergraduate and graduate students as well as new researchers, developers, and practitioners in this field can easily trust and grasp the concepts, and learn them effectively. It has been written to reduce the mathematical complexity and help the vast majority of readers to understand the topics and get interested in the field. This book consists of four parts, with the total of 14 chapters. The first part mainly focuses on the topics that are needed to help analyze and understand data and big data. The second part covers the topics that can explain the systems required for processing big data. The third part presents the topics required to understand and select machine learning techniques to classify big data. Finally, the fourth part concentrates on the topics that explain the scaling-up machine learning, an important solution for modern big data problems.

Big Data Computing - Rajendra Akerkar 2013-12-05

Due to market forces and technological evolution, Big Data computing is developing at an increasing rate. A wide variety of novel approaches and tools have emerged to tackle the challenges of Big Data, creating both more opportunities and more challenges for students and professionals in the field of data computation and analysis. Presenting a mix of industry cases and theory, Big Data Computing discusses the technical and practical issues related to Big Data in intelligent information management. Emphasizing the adoption and diffusion of Big Data tools and technologies in industry, the book introduces a broad range of Big Data concepts, tools, and techniques. It covers a wide range of research, and provides comparisons between state-of-the-art approaches. Comprised of five sections, the book focuses on: What Big Data is and why it is important Semantic technologies Tools and methods Business and economic perspectives Big Data applications across industries

Principles of Data Science - Sinan Ozdemir 2016-12-16

Learn the techniques and math you need to start making sense of your data About This Book Enhance your knowledge of coding with data science theory for practical insight into data science and analysis More than just a math class, learn how to perform real-world data science tasks with R and Python Create actionable insights and transform raw data into tangible value Who This Book Is For You should be fairly well acquainted with basic algebra and should feel comfortable reading snippets of R/Python as well as pseudo code. You should have the urge to learn and apply the techniques put forth in this book on either your own data sets or those provided to you. If you have the basic math skills but want to apply them in data science or you have good programming skills but lack math, then this book is for you. What You Will Learn Get to know the five most important steps of data science Use your data intelligently and learn how to handle it with care Bridge the gap between mathematics and programming Learn about probability, calculus, and how to use statistical models to control and clean your data and drive actionable results Build and evaluate baseline machine learning models Explore the most effective metrics to determine the success of your machine learning models Create data visualizations that communicate actionable insights Read and apply machine learning concepts to your problems and make actual predictions In Detail Need to turn your skills at programming into effective data science skills? Principles of Data Science is created to help you join the dots between mathematics, programming, and business analysis. With this book, you'll feel confident about asking—and answering—complex and sophisticated questions of your data to move from abstract and raw statistics to actionable ideas. With a unique approach that bridges the gap between mathematics and computer science, this books takes you through the entire data science pipeline. Beginning with cleaning and preparing data, and effective data mining strategies and techniques, you'll move on to build a comprehensive picture of how every piece of the data science puzzle fits together. Learn the fundamentals of computational mathematics and statistics, as well as some pseudocode being used today by data scientists and analysts. You'll get to grips with machine learning, discover the statistical models that help you take control and navigate even the densest datasets, and find out how to create powerful visualizations that communicate what your data means. Style and approach This is an easy-to-understand and accessible tutorial. It is a step-by-step guide with use cases, examples, and illustrations to get you well-versed with the concepts of data science. Along with explaining the fundamentals, the book will also introduce you to slightly advanced concepts later on and will help you implement these techniques in the real world.

Data Pipelines Pocket Reference - James Densmore 2021-02-10

Data pipelines are the foundation for success in data analytics. Moving data from numerous diverse sources and transforming it to provide context is the difference between having data and actually gaining value from it. This pocket reference defines data pipelines and explains how they work in today's modern data stack. You'll learn common considerations and key decision points when implementing pipelines, such as batch versus streaming data ingestion and build versus buy. This book addresses the most common decisions made by data professionals and discusses foundational concepts that apply to open source frameworks, commercial products, and homegrown solutions. You'll learn: What a data pipeline is and how it works How data is moved and processed on modern data infrastructure, including cloud platforms Common tools and products used by data engineers to build pipelines How pipelines support analytics and reporting needs Considerations for pipeline maintenance, testing, and alerting

Hadoop in Practice - Alex Holmes 2014-09-29

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

Principles of Big Data - Jules J. Berman 2013-05-20

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

Practical Enterprise Data Lake Insights - Saurabh Gupta 2018-07-29

Use this practical guide to successfully handle the challenges encountered when designing an enterprise data lake and learn industry best practices to resolve issues. When designing an enterprise data lake you often hit a roadblock when you must leave the comfort of the relational world and learn the nuances of handling non-relational data. Starting from sourcing data into the Hadoop ecosystem, you will go through stages that can bring up tough questions such as data processing, data querying, and security. Concepts such as change data capture and data streaming are covered. The book takes an end-to-end solution approach in a data lake environment that includes data security, high availability, data processing, data streaming, and more. Each chapter includes application of a concept, code snippets, and use case demonstrations to provide you with a practical approach. You will learn the concept, scope, application, and starting point. What You'll Learn Get to know data lake architecture and design principles Implement data capture and streaming strategies Implement data processing strategies in Hadoop Understand the data lake security framework and availability model Who This Book Is For Big data architects and solution architects