Big Data Principles Practices Scalable

Thank you enormously much for downloading **Big Data Principles Practices Scalable** .Most likely you have knowledge that, people have look numerous times for their favorite books similar to this Big Data Principles Practices Scalable , but end taking place in harmful downloads.

Rather than enjoying a good ebook similar to a cup of coffee in the afternoon, otherwise they juggled behind some harmful virus inside their computer. **Big Data Principles Practices Scalable** is nearby in our digital library an online access to it is set as public as a result you can download it instantly. Our digital library saves in merged countries, allowing you to acquire the most less latency epoch to download any of our books subsequent to this one. Merely said, the Big Data Principles Practices Scalable is universally compatible with any devices to read.

Web Services: Concepts, Methodologies, Tools, and Applications - Management Association, Information Resources 2018-12-07
Web service technologies are redefining the way that large and small companies are doing business and exchanging information. Due to the critical need for furthering automation, engagement, and efficiency, systems and workflows are becoming increasingly more web-based. Web Services: Concepts, Methodologies, Tools, and Applications is an innovative reference source that examines relevant theoretical frameworks, current practice guidelines, industry standards and

standardization, and the latest empirical research findings in web services. Highlighting a range of topics such as cloud computing, quality of service, and semantic web, this multi-volume book is designed for computer engineers, IT specialists, software designers, professionals, researchers, and upper-level students interested in web services architecture, frameworks, and security.

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

Solving Large Scale Learning Tasks. Challenges and Algorithms - Stefan Michaelis 2016-07-02

In celebration of Prof. Morik's 60th birthday, this Festschrift covers research areas that Prof. Morik worked in and presents various researchers with whom she collaborated. The 23 refereed articles in this Festschrift volume provide challenges and solutions from theoreticians and practitioners on data preprocessing, modeling, learning, and evaluation. Topics include data-mining and machine-learning algorithms, feature selection and feature generation, optimization as well as efficiency of energy and communication.

Understanding Distributed Systems - Roberto Vitillo 2021

Learning to build distributed systems is hard, especially if they are large scale. It's not that there is a lack of information out there. You can find academic papers, engineering blogs, and even books on the subject. The problem is that the available information is spread out all over the place, and if you were to put it on a spectrum from theory to practice, you would find a lot of material at the two ends, but not much in the middle. That is why I decided to write a book to teach the

fundamentals of distributed systems so that you don't have to spend countless hours scratching your head to understand how everything fits together. This is the guide I wished existed when I first started out, and it's based on my experience building large distributed systems that scale to millions of requests per second and billions of devices. If you develop the back-end of web or mobile applications (or would like to!), this book is for you. When building distributed systems, you need to be familiar with the network stack, data consistency models, scalability and reliability patterns, and much more. Although you can build applications without knowing any of that, you will end up spending hours debugging and re-designing their architecture, learning lessons that you could have acquired in a much faster and less painful way.

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.

Designing Distributed Systems - Brendan Burns 2018-02-20

Without established design patterns to guide them, developers have had to build distributed systems from scratch, and most of these systems are very unique indeed. Today, the increasing use of containers has paved the way for core distributed system patterns and reusable containerized components. This practical guide presents a collection of repeatable, generic patterns to help make the development of reliable distributed systems far more approachable and efficient. Author Brendan Burns—Director of Engineering at Microsoft Azure—demonstrates how you can adapt existing software design patterns for designing and building reliable distributed applications.

Systems engineers and application developers will learn how these long-established patterns provide a common language and framework for dramatically increasing the quality of your system. Understand how patterns and reusable components enable the rapid development of reliable distributed systems Use the side-car, adapter, and ambassador patterns to split your application into a group of containers on a single machine Explore loosely coupled multi-node distributed patterns for replication, scaling, and communication between the components Learn distributed system patterns for large-scale batch data processing covering work-queues, event-based processing, and coordinated workflows

Architecting High Performing, Scalable and Available Enterprise Web Applications - Shailesh Kumar Shivakumar 2014-10-29 Architecting High Performing, Scalable and Available Enterprise Web Applications provides in-depth insights into techniques for achieving desired scalability, availability and performance quality goals for enterprise web applications. The book provides an integrated 360-degree view of achieving and maintaining these attributes through practical, proven patterns, novel models, best practices, performance strategies, and continuous improvement methodologies and case studies. The author shares his years of experience in application security, enterprise application testing, caching techniques, production operations and maintenance, and efficient project management techniques. Delivers holistic view of scalability, availability and security, caching, testing and project management Includes patterns and frameworks that are illustrated with end-to-end case studies Offers tips and troubleshooting methods for enterprise application testing, security, caching, production operations and project management Exploration of

synergies between techniques and methodologies to achieve end-to-end availability, scalability, performance and security quality attributes 360-degree viewpoint approach for achieving overall quality Practitioner viewpoint on proven patterns, techniques, methodologies, models and best practices. Bulleted summary and tabular representation of concepts for effective understanding Production operations and troubleshooting tips

Mastering Azure Analytics - Zoiner Tejada 2017-04-06 Helps users understand the breadth of Azure services by organizing them into a reference framework they can use when crafting their own big-data analytics solution.

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

Scalability Rules - Martin L. Abbott 2011-05-04

50 Powerful, Easy-to-Use Rules for Supporting Hypergrowth in Any Environment Scalability Rules is the easy-to-use scalability primer and reference for every architect, developer, web professional, and manager. Authors Martin L. Abbott and Michael T. Fisher have helped scale more than 200 hypergrowth Internet sites through their consulting practice. Now, drawing on their unsurpassed experience, they present 50 clear, proven scalability rules—and practical guidance for applying them. Abbott and Fisher transform scalability from a "black art" to a set of realistic, technology-agnostic best practices for supporting hypergrowth in nearly any environment, including both frontend and backend systems. For architects, they offer powerful new insights for creating and evaluating designs. For developers, they share specific techniques for handling everything from databases to state. For managers, they provide invaluable help in goal-setting, decision-making, and interacting with technical teams. Whatever your role, you'll find practical risk/benefit guidance for setting priorities—and getting maximum "bang for the buck." • Simplifying architectures and avoiding "over-engineering" • Scaling via cloning, replication, separating functionality, and splitting data sets • Scaling out, not up • Getting more out of databases without compromising scalability • Avoiding unnecessary redirects and redundant double-checking • Using caches and content delivery networks more aggressively, without introducing unacceptable complexity • Designing for fault tolerance, graceful failure, and easy rollback • Striving for statelessness when you can; efficiently handling state when you must • Effectively utilizing asynchronous communication • Learning quickly from mistakes, and much more

Data Management at Scale - Piethein Strengholt 2020-07-29 As data management and integration continue to evolve rapidly, storing all your data in one place, such as a data warehouse, is no longer scalable. In the very near future, data will need to be distributed and available for several technological solutions. With this practical book, you'll learnhow to migrate your enterprise from a complex and tightly coupled data landscape to a more flexible architecture ready for the modern world of data consumption. Executives, data architects, analytics teams, and compliance and governance staff will learn how to build a modern scalable data landscape using the Scaled Architecture, which you can introduce incrementally without a large upfront investment. Author Piethein Strengholt provides blueprints, principles, observations, best practices, and patterns to get you up to speed. Examine data management trends, including technological developments, regulatory requirements, and privacy concerns Go deep into the Scaled Architecture and learn how the pieces fit together Explore data governance and data security, master data management, self-service data marketplaces, and the importance of metadata

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

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

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

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.

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

Microservice Architecture - Irakli Nadareishvili 2016-07-18
Have you heard about the tremendous success Amazon and Netflix have had by switching to a microservice architecture? Are you wondering how this can benefit your company? Or are you skeptical about how it might work? If you've answered yes to any of these questions, this practical book will benefit you. You'll learn how to take advantage of the microservice architectural style for building systems, and learn from the experiences of others to adopt and execute this approach most successfully.

Data Lake for Enterprises - Tomcy John 2017-05-31
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.

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 - 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

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

Frontiers in Massive Data Analysis - National Research Council 2013-09-03

Data mining of massive data sets is transforming the way we think about crisis response, marketing, entertainment, cybersecurity and national intelligence. Collections of documents, images, videos, and networks are being thought of not merely as bit strings to be stored, indexed, and retrieved, but as potential sources of discovery and knowledge, requiring sophisticated analysis techniques that go far beyond classical indexing and keyword counting, aiming to find relational and semantic interpretations of the phenomena underlying the data. Frontiers in Massive Data Analysis examines the frontier of analyzing massive amounts of data, whether in a static database or streaming through a system. Data at that scale--terabytes and petabytes--is increasingly common in science (e.g., particle physics, remote sensing, genomics), Internet commerce, business analytics, national security, communications, and elsewhere. The tools that work to infer knowledge from data at smaller scales do not necessarily work, or work well, at such massive scale. New tools, skills, and approaches are necessary, and this report identifies many of them, plus promising research directions to explore. Frontiers in Massive Data Analysis discusses pitfalls in trying to infer knowledge from massive data, and it characterizes seven major classes of computation that are common in the analysis of massive data.

Overall, this report illustrates the cross-disciplinary knowledge--from computer science, statistics, machine learning, and application disciplines--that must be brought to bear to make useful inferences from massive data.

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 - 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.

High-Performance Modelling and Simulation for Big Data Applications - Joanna Kołodziej 2019-03-25

This open access book was prepared as a Final Publication of the COST Action IC1406 “High-Performance Modelling and Simulation for Big Data Applications (cHiPSet)“ project. Long considered important pillars of the scientific method, Modelling and Simulation have evolved from traditional discrete numerical methods to complex data-intensive continuous analytical optimisations. Resolution, scale, and accuracy have become essential to predict and analyse natural and complex systems in science and engineering. When their level of abstraction raises to have a better discernment of the domain at hand, their representation gets increasingly demanding for computational and data resources. On the other hand, High Performance Computing typically entails the effective use of parallel and distributed processing units coupled with efficient storage, communication and visualisation systems to underpin complex data-intensive applications in distinct scientific and technical domains. It is then arguably required to have a seamless interaction of High Performance Computing with Modelling and Simulation in order to store, compute, analyse, and visualise large data sets in science and engineering. Funded by the European Commission, cHiPSet has provided a dynamic trans-European forum for their members and distinguished guests to openly discuss novel perspectives and topics of interests for these two communities. This cHiPSet compendium presents a set of selected case studies related to healthcare, biological data, computational advertising, multimedia, finance, bioinformatics, and telecommunications.

Building a Scalable Data Warehouse with Data Vault 2.0 - Dan Linstedt 2015-09-15

The Data Vault was invented by Dan Linstedt at the U.S. Department of Defense, and the standard has been successfully applied to data warehousing projects at organizations of different sizes, from small to large-size corporations. Due to its simplified design, which is adapted from nature, the Data Vault 2.0 standard helps prevent typical data warehousing failures. "Building a Scalable Data Warehouse" covers everything one needs to know to create a scalable data warehouse end to end, including a presentation of the Data Vault modeling technique, which provides the foundations to create a technical data warehouse layer. The book discusses how to build the data warehouse incrementally using the agile Data Vault 2.0 methodology. In addition, readers will learn how to create the input layer (the stage layer) and the presentation layer (data mart) of the Data Vault 2.0 architecture including implementation best practices. Drawing upon years of practical experience and using numerous examples and an easy to understand framework, Dan Linstedt and Michael Olschimke discuss: How to load each layer using SQL Server Integration Services (SSIS), including automation of the Data Vault loading processes. Important data warehouse technologies and practices. Data Quality Services (DQS) and Master Data Services (MDS) in the context of the Data Vault architecture. Provides a complete introduction to data warehousing, applications, and the business context so readers can get-up and running fast Explains theoretical concepts and provides hands-on instruction on how to build and implement a data warehouse Demystifies data vault modeling with beginning, intermediate, and advanced techniques Discusses the advantages of the data vault approach over other techniques, also including the latest updates to Data Vault 2.0 and multiple improvements to

Data Vault 1.0

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 Mesh - Zhamak Dehghani 2022-03-08

We're at an inflection point in data, where our data management solutions no longer match the complexity of organizations, the proliferation of data sources, and the scope of our aspirations to get value from data with AI and analytics. In this practical book, author Zhamak Dehghani introduces data mesh, a decentralized sociotechnical paradigm drawn from modern distributed architecture that provides a new approach to sourcing, sharing, accessing,

and managing analytical data at scale. Dehghani guides practitioners, architects, technical leaders, and decision makers on their journey from traditional big data architecture to a distributed and multidimensional approach to analytical data management. Data mesh treats data as a product, considers domains as a primary concern, applies platform thinking to create self-serve data infrastructure, and introduces a federated computational model of data governance. Get a complete introduction to data mesh principles and its constituents Design a data mesh architecture Guide a data mesh strategy and execution Navigate organizational design to a decentralized data ownership model Move beyond traditional data warehouses and lakes to a distributed data mesh

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

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

New Horizons for a Data-Driven Economy - José María Cavanillas 2016-04-04

In this book readers will find technological discussions on the existing and emerging technologies across the different stages of the big data value chain. They will learn about legal aspects of big data, the social impact, and about education needs and requirements. And they will discover the business perspective and how big data technology can be exploited to deliver value within different sectors of the economy. The book is structured in four parts: Part I "The Big Data Opportunity" explores the value potential of big data with a particular focus on the European context. It also describes the legal, business and social dimensions that

need to be addressed, and briefly introduces the European Commission's BIG project. Part II "The Big Data Value Chain" details the complete big data lifecycle from a technical point of view, ranging from data acquisition, analysis, curation and storage, to data usage and exploitation. Next, Part III "Usage and Exploitation of Big Data" illustrates the value creation possibilities of big data applications in various sectors, including industry, healthcare, finance, energy, media and public services. Finally, Part IV "A Roadmap for Big Data Research" identifies and prioritizes the cross-sectorial requirements for big data research, and outlines the most urgent and challenging technological, economic, political and societal issues for big data in Europe. This compendium summarizes more than two years of work performed by a leading group of major European research centers and industries in the context of the BIG project. It brings together research findings, forecasts and estimates related to this challenging technological context that is becoming the major axis of the new digitally transformed business environment.

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

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

Exploring Enterprise Service Bus in the Service-Oriented Architecture Paradigm - Bhadoria, Robin Singh 2017-02-14

Web browsing would not be what it is today without the use

of Service-Oriented Architecture (SOA). Although much has been written about SOA methodology, this emerging platform is continuously under development. Exploring Enterprise Service Bus in the Service-Oriented Architecture Paradigm is a detailed reference source that examines current aspects and research methodologies that enable enterprise service bus to unify and connect services efficiently on a common platform. Featuring relevant topics such as SOA reference architecture, grid computing applications, complex event computing, and java business integration, this is an ideal resource for all practitioners, academicians, graduate students, and researchers interested in the discoveries on the relationship that Service-Oriented architecture and enterprise service bus share.

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.

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.

Ethics of Big Data - Kord Davis 2012-09-13

What are your organization's policies for generating and using huge datasets full of personal information? This book examines ethical questions raised by the big data phenomenon, and explains why enterprises need to reconsider business decisions concerning privacy and identity. Authors Kord Davis and Doug Patterson provide methods and techniques to help your business engage in a transparent and productive ethical inquiry into your current data practices. Both individuals and organizations have legitimate interests in understanding how data is handled. Your use of data can directly affect brand quality and revenue—as Target, Apple, Netflix, and dozens of other companies have discovered. With this book, you'll learn how to align your actions with explicit company values and preserve the trust of customers, partners, and stakeholders. Review your data-handling practices and examine whether they reflect core organizational values Express coherent and consistent positions on your organization's use of big data Define tactical plans to close gaps between values and practices—and discover how to maintain alignment as conditions change over time Maintain a balance between the benefits of innovation and the risks of unintended consequences

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.

The Art of Scalability - Martin L. Abbott 2015-05-23 The Comprehensive, Proven Approach to IT Scalability–Updated with New Strategies, Technologies, and Case Studies In The Art of Scalability, Second Edition, leading scalability consultants Martin L. Abbott and Michael T. Fisher cover everything you need to know to smoothly scale products and services for any requirement. This extensively revised edition reflects new technologies, strategies, and lessons, as well as new case studies from the authors' pioneering consulting practice, AKF Partners. Writing for technical and nontechnical decision-makers, Abbott and Fisher cover everything that impacts scalability, including architecture, process, people, organization, and technology. Their insights and

recommendations reflect more than thirty years of experience at companies ranging from eBay to Visa, and Salesforce.com to Apple. You'll find updated strategies for structuring organizations to maximize agility and scalability, as well as new insights into the cloud (IaaS/PaaS) transition, NoSQL, DevOps, business metrics, and more. Using this guide's tools and advice, you can systematically clear away obstacles to scalability–and achieve unprecedented IT and business performance. Coverage includes • Why scalability problems start with organizations and people, not technology, and what to do about it • Actionable lessons from real successes and failures • Staffing, structuring, and leading the agile, scalable organization • Scaling processes for hyper-growth environments • Architecting scalability: proprietary models for clarifying needs and making choices–including 15 key success principles • Emerging technologies and challenges: data cost, datacenter planning, cloud evolution, and customer-aligned monitoring • Measuring availability, capacity, load, and performance

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