Big Data Principles Practices Scalable

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

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*

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*

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.*

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.

Hadoop in Practice

The Enterprise Big Data Lake

The Complete Guide to Large-Scale Analysis and Modeling

Best Practices for Designing, Implementing, and Maintaining Systems

Principles of Database Management

Work with massive datasets to design data models and automate data pipelines using Python

Understanding the real-time pipeline

Integrate big data into business to drive competitive advantage and sustainable success Big Data MBA brings insight and expertise to leveraging big data in business so you can harness the power of analytics and gain a true business advantage. Based on a practical framework with supporting methodology and hands-on exercises, this book helps identify where and how big data can help you transform your business. You'll learn how to exploit new sources of customer, product, and operational data, coupled with advanced analytics and data science, to optimize key processes, uncover monetization opportunities, and create new sources of competitive differentiation. The discussion includes guidelines for operationalizing analytics, optimal organizational structure, and using analytic insights throughout your organization's user experience to customers and front-end employees alike. You'll learn to “think like a data scientist” as you build upon the decisions your business is trying to make, the hypotheses you need to test, and the predictions you need to produce. Business stakeholders no longer need to relinquish control of data and analytics to IT. In fact, they must champion the organization's data collection and analysis efforts. This book is a primer on the business approach to analytics, providing the practical understanding you need to convert data into opportunity. Understand where and how to leverage big data Integrate analytics into everyday operations Structure your organization to drive analytic insights Optimize processes, uncover opportunities, and stand out from the rest Help business stakeholders to “think like a data scientist” Understand appropriate business application of different analytic techniques If you want data to transform your business, you need to know how to put it to use. Big Data MBA shows you how to implement big data and analytics to make better decisions.

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.

Connects fundamental mathematical theory with real-world problems, through efficient and scalable optimization algorithms.

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

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

Principles, Computation, and Applications

Patterns and Paradigms for Scalable, Reliable Services

Building a Scalable Data Warehouse with Data Vault 2.0

Designing Distributed Systems

Best Practices for Large Enterprises

Building Standardized Systems Across an Engineering Organization

Microservice Architecture

This book provides an introduction to the mathematical and algorithmic foundations of data science, including machine learning, high-dimensional geometry, and analysis of large networks. Topics include the counterintuitive nature of data in high dimensions, important linear algebraic techniques such as singular value decomposition, the theory of random walks and Markov chains, the fundamentals of and important algorithms for machine learning, algorithms and analysis for clustering, probabilistic models for large networks, representation learning including topic modelling and non-negative matrix factorization, wavelets and compressed sensing. Important probabilistic techniques are developed including the law of large numbers, tail inequalities, analysis of random projections, generalization guarantees in machine learning, and moment methods for analysis of phase transitions in large random graphs. Additionally, important structural and complexity measures are discussed such as matrix norms and VC-dimension. This book is suitable for both undergraduate and graduate courses in the design and analysis of algorithms for data.

“Companies have been implementing large agile projects for a number of years, but the ‘stigma’ of ‘agile only works for small projects’ continues to be a frequent barrier for newcomers and a rallying cry for agile critics. What has been missing from the agile literature is a solid, practical book on the specifics of developing large projects in an agile way. Dean Leffingwell’s book Scaling Software Agility fills this gap admirably. It offers a practical guide to large project issues such as architecture, requirements development, multi-level release planning, and team organization. Leffingwell’s book is a necessary guide for large projects and large organizations making the transition to agile development." —Jim Highsmith, director, Agile Practice, Cutter Consortium, author of Agile Project Management "There's tension between building software fast and delivering software that lasts, between being ultra-responsive to changes in the market and maintaining a degree of stability. In his latest work, Scaling Software Agility, Dean Leffingwell shows how to achieve a pragmatic balance among these forces. Leffingwell's observations of the problem, his advice on the solution, and his description of the resulting best practices come from experience: he's been there, done that, and has seen what's worked." —Grady Booch, IBM Fellow Agile development practices, while still controversial in some circles, offer undeniable benefits: faster time to market, better responsiveness to changing customer requirements, and higher quality. However, agile practices have been defined and recommended primarily to small teams. In Scaling Software Agility, Dean Leffingwell describes how agile methods can be applied to enterprise-class development. Part I provides an overview of the most common and effective agile methods. Part II describes seven best practices of agility that natively scale to the enterprise level. Part III describes an additional set of seven organizational capabilities that companies can master to achieve the full benefits of software agility on an enterprise scale. This book is invaluable to software developers, testers and QA personnel, managers and team leads, as well as to executives of software organizations whose objective is to increase the quality and productivity of the software development process but who are faced with all the challenges of developing software on an enterprise scale.

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

Fully updated! Fifty Powerful, Easy-to-Use Rules for Supporting Hyper Growth “Whether you’re taking on a role as a technology leader in a new company or you simply want to make great technology decisions, Scalability Rules will be the go-to resource on your bookshelf.” –Chad Dickerson, CTO, Etsy Scalability Rules, Second Edition, is the easy-to-use scalability primer and reference for every architect, developer, network/software engineer, web professional, and manager. Authors Martin L. Abbott and Michael T. Fisher have helped scale hundreds of high-growth companies and thousands of systems. Drawing on their immense experience, they present 50 up-to-the-minute technical best practices for supporting hyper growth practically anywhere. Fully updated to reflect new technical trends and experiences, this edition is even easier to read, understand, and apply. Abbott and Fisher have also added powerful "stories behind the rules": actual experiences and case studies from CTOs and technology executives at Etsy, NASDAQ, Salesforce, Shutterfly, Chegg, Warby Parker, Twitter, and other scalability pioneers. Architects will find powerful technology-agnostic insights for creating and evaluating designs. Developers will discover specific techniques for handling everything from databases to state. Managers will get invaluable help in setting goals, making decisions, and interacting with technical teams. Whatever your role, you'll find practical risk/benefit guidance for setting priorities, translating plans into action, and gaining maximum scalability at minimum cost. You'll learn how to Simplify architectures and avoid "over-engineering" Design scale into your solution, so you can scale on a just-in-time basis Make the most of cloning and replication Separate functionality and split data sets Scale out, not up Get more out of databases without compromising scalability Eliminate unnecessary redirects and redundant double-checking Use caches and CDNs more aggressively, without unacceptable complexity Design for fault tolerance, graceful failure, and easy rollback Emphasize statelessness, and efficiently handle state when you must Effectively utilize asynchronous communication Learn from your own mistakes and others' high-profile failures Prioritize your actions to get the biggest “bang for the buck”

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.

Thinking with Examples for Effective Learning

The Practical Guide to Storing, Managing and Analyzing Big and Small Data

Principles and Practice

Principles and Paradigms

Knowledge Graphs and Big Data Processing

Machine Learning Models and Algorithms for Big Data Classification

Selected Results of the COST Action IC1406 cHiPSet

Summary Streaming Data introduces the concepts and requirements of streaming and real-time data systems. The book is an idea-rich tutorial that teaches you to think about how to efficiently interact with fast-flowing data. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology As humans, we're constantly filtering and deciphering the information streaming

toward us. In the same way, streaming data applications can accomplish amazing tasks like reading live location data to recommend nearby services, tracking faults with machinery in real time, and sending digital receipts before your customers leave the shop. Recent advances in streaming data technology and techniques make it possible for any developer to build these applications if they have the right mindset. This book will let you join them. About the Book Streaming Data is an idea-rich tutorial that teaches you to think about efficiently interacting with fast-flowing data. Through relevant examples and illustrated use cases, you'll explore designs for applications that read, analyze, share, and store streaming data. Along the way, you'll discover the roles of key technologies like Spark, Storm, Kafka, Flink, RabbitMQ, and more. This book offers the perfect balance between big-picture thinking and implementation details. What's Inside The right way to collect real-time data Architecting a streaming pipeline Analyzing the data Which technologies to use and when About the Reader Written for developers familiar with relational database concepts. No experience with streaming or real-time applications required.

About the Author Andrew Psaltis is a software engineer focused on massively scalable real-time analytics. Table of Contents PART 1 - A NEW HOLISTIC APPROACH Introducing streaming data Getting data from clients: data ingestion Transporting the data from collection tier: decoupling the data pipeline Analyzing streaming data Algorithms for data analysis Storing the analyzed or collected data Making the data available Consumer device capabilities and limitations accessing the data PART 2 - TAKING IT REAL WORLD Analyzing Meetup RSVPs in real time

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

“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 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. 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.
Microservices can have a positive impact on your enterprise—just ask Amazon and Netflix—but you can fall into many traps if you don’t approach them in the right way. This practical guide covers the entire microservices landscape, including the principles, technologies, and methodologies of this unique, modular style of system building. You’ll learn about the

experiences of organizations around the globe that have successfully adopted microservices. In three parts, this book explains how these services work and what it means to build an application the Microservices Way. You'll explore a design-based approach to microservice architecture with guidance for implementing various elements. And you'll get a set of recipes and practices for meeting practical, organizational, and cultural challenges to microservice adoption. Learn how microservices can help you drive business objectives Examine the principles, practices, and culture that define microservice architectures Explore a model for creating complex systems and a design process for building a microservice architecture Learn the fundamental design concepts for individual microservices Delve into the operational elements of a microservices architecture, including containers and service discovery Discover how to handle the challenges of introducing microservice architecture in your organization

Scaling Software Agility

Concepts, Drivers & Techniques

Building Secure and Reliable Systems

50 Principles for Scaling Web Sites

Production-Ready Microservices

High-Dimensional Data Analysis with Low-Dimensional Models

Big Data

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 Introductory, theory-practice balanced text teaching the fundamentals of databases to advanced undergraduates or graduate students in information systems or computer science. 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 DataPrinciples and Best Practices of Scalable Realtime Data SystemsManning Publications Company

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.

Principles of Big Data

New Horizons for a Data-Driven Economy

Big Data Fundamentals

Big Data Management

A guide to building proficiency in tools and systems used by leading big data experts

Data Management at Scale

Data Pipelines Pocket Reference

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

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.

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.

From the Foreword: "Big Data Management and Processing is [a] state-of-the-art book that deals with a wide range of topical themes in the field of Big Data. The book, which probes many issues related to this exciting and rapidly growing field, covers processing, management, analytics, and applications... [It] is a very valuable addition to the literature. It will serve as a source of up-to-date research in this continuously developing area. The book also provides an opportunity for researchers to explore the use of advanced computing technologies and their impact on enhancing our capabilities to conduct more sophisticated studies." ---Sartaj Sahni, University of Florida, USA "Big

Data Management and Processing covers the latest Big Data research results in processing, analytics, management and applications. Both fundamental insights and representative applications are provided. This book is a timely and valuable resource for students, researchers and seasoned practitioners in Big Data fields. --Hai Jin, Huazhong University of Science and Technology, China Big Data Management and Processing explores a range of big data related issues and their impact on the design of new computing systems. The twenty-one chapters were carefully selected and feature contributions from several outstanding researchers. The book endeavors to strike a balance between theoretical and practical coverage of innovative problem solving techniques for a range of platforms. It serves as a repository of paradigms, technologies, and applications that target different facets of big data computing systems. The first part of the book explores energy and resource management issues, as well as legal compliance and quality management for Big Data. It covers In-Memory computing and

In-Memory data grids, as well as co-scheduling for high performance computing applications. The second part of the book includes comprehensive coverage of Hadoop and Spark, along with security, privacy, and trust challenges and solutions. The latter part of the book covers mining and clustering in Big Data, and includes applications in genomics, hospital big data processing, and vehicular cloud computing. The book also analyzes funding for Big Data projects.

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

Ethics of Big Data

High-Performance Modelling and Simulation for Big Data Applications

Preparing, Sharing, and Analyzing Complex Information

Principles and Best Practices of Scalable Realtime Data Systems

Data Governance Principles for Big Data Analytics

Big Data Architect’s Handbook

A Roadmap for Usage and Exploitation of Big Data in Europe

In the race to compete in today’s fast-moving markets, large enterprises are busy adopting new technologies for creating new products, processes, and business models. But one obstacle on the road to digital transformation is placing too much emphasis on technology, and not enough on the types of processes technology enables. What if different lines of business could build their own services and applications—and decision-making was distributed rather than centralized? This report explores the concept of a digital business platform as a way of empowering individual business sectors to act on data in real time. Much innovation in a digital enterprise will increasingly happen at the edge, whether it involves business users (from marketers to data scientists) or IoT devices. To facilitate the process, your core IT team can provide these sectors with the digital tools they need to innovate quickly. This report explores: Key cultural and organizational changes for developing business capabilities through cross-functional product teams A platform for integrating applications, data sources, business partners, clients, mobile apps, social networks, and IoT devices

Creating internal API programs for building innovative edge services in low-code or no-code environments Tools including Integration Platform as a Service, Application Platform as a Service, and Integration Software as a Service The challenge of integrating microservices and serverless architectures Event-driven architectures for processing and reacting to events in real time You'll also learn about a complete pervasive integration solution as a core component of a digital business platform to serve every audience in your organization.

A comprehensive end-to-end guide that gives hands-on practice in big data and Artificial Intelligence Key Features Learn to build and run a big data application with sample code Explore examples to implement activities that a big data architect performs Use Machine Learning and AI for structured and unstructured data Book Description The big data architects are the "masters" of data, and hold high value in today's market. Handling big data, be it of good or bad quality, is not an easy task. The prime job for any big data architect is to build an end-to-end big data solution that integrates data from different sources and analyzes it to find useful, hidden insights. Big Data Architect's Handbook takes you through

developing a complete, end-to-end big data pipeline, which will lay the foundation for you and provide the necessary knowledge required to be an architect in big data. Right from understanding the design considerations to implementing a solid, efficient, and scalable data pipeline, this book walks you through all the essential aspects of big data. It also gives you an overview of how you can leverage the power of various big data tools such as Apache Hadoop and ElasticSearch in order to bring them together and build an efficient big data solution. By the end of this book, you will be able to build your own design system which integrates, maintains, visualizes, and monitors your data. In addition, you will have a smooth design flow in each process, putting insights in action. What you will learn Learn Hadoop Ecosystem and Apache projects Understand, compare NoSQL database and essential software architecture Cloud infrastructure design considerations for big data Explore application scenario of big data tools for daily activities Learn to analyze and visualize results to uncover valuable insights Build and run a big data application with sample code from end to end Apply Machine Learning and AI to perform big data intelligence Practice the daily activities performed by big data architects Who

this book is for Big Data Architect's Handbook is for you if you are an aspiring data professional, developer, or IT enthusiast who aims to be an all-round architect in big data. This book is your one-stop solution to enhance your knowledge and carry out easy to complex activities required to become a big data architect.

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
Can a system be considered truly reliable if it isn't fundamentally secure? Or can it be considered secure if it's unreliable? Security is crucial to the design and operation of scalable systems in production, as it plays an important part in product quality, performance, and availability. In this book, experts from Google share best practices to help your organization design scalable and reliable systems that are fundamentally secure. Two previous O’Reilly books from Google—Site Reliability Engineering and The Site Reliability Workbook—demonstrated how and why a commitment to the entire service lifecycle enables organizations to successfully build, deploy, monitor, and maintain software systems. In this latest guide, the authors offer insights into system design, implementation, and maintenance from practitioners who specialize in security and reliability. They also discuss how building and adopting their recommended best practices requires a culture that’s supportive of such change. You’ll learn about secure and reliable systems through: Design strategies Recommendations for coding,

testing, and debugging practices Strategies to prepare for, respond to, and recover from incidents Cultural best practices that help teams across your organization collaborate effectively

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 Management and Processing

Information Governance Principles and Practices for a Big Data Landscape

Foundations of Data Science

Mastering Spark with R

Architecting High Performing, Scalable and Available Enterprise Web Applications

Designing Data-Intensive Applications

97 Things Every Data Engineer Should Know

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*

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.*

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*

One of the biggest challenges for organizations that have adopted microservice architecture is the lack of architectural, operational, and organizational standardization. After splitting a monolithic application or building a microservice ecosystem from scratch, many engineers are left wondering what’s next. In this practical book, author Susan Fowler presents a set of microservice standards in depth, drawing from her experience standardizing over a thousand microservices at Uber. You’ll learn how to design microservices that are stable, reliable, scalable, fault tolerant, performant, monitored, documented, and prepared for any catastrophe. Explore production-readiness standards, including: Stability and Reliability: develop, deploy, introduce, and deprecate microservices; protect against *dependency failures Scalability and Performance: learn essential components for achieving greater microservice efficiency Fault Tolerance and Catastrophe Preparedness: ensure availability by actively pushing microservices to fail in real time Monitoring: learn how to monitor, log, and display key metrics; establish alerting and on-call procedures Documentation and Understanding: mitigate tradeoffs that come with microservice adoption, including organizational sprawl and technical debt*

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

Streaming Data
Aligning Principles, Practices, and Culture
Balancing Risk and Innovation
Big Data MBA
Big Data For Dummies
Data Engineering with Python

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

The Big Ideas Behind Reliable, Scalable, and Maintainable Systems

Driving Business Strategies with Data Science

Data Privacy

Delivering the Promise of Big Data and Data Science

Frontiers in Massive Data Analysis

Real-World Machine Learning

Scalability Rules