BIG DATA HADOOP E HIVE Che Cosa Sono In Breve E A Cosa Servono

Yeah, reviewing a book **BIG DATA HADOOP E HIVE Che Cosa Sono In Breve E A Cosa Servono** could ensue your close links listings. This is just one of the solutions for you to be successful. As understood, achievement does not suggest that you have wonderful points.

Comprehending as skillfully as treaty even more than supplementary will have enough money each success. adjacent to, the pronouncement as capably as insight of this BIG DATA HADOOP E HIVE Che Cosa Sono In Breve E A Cosa Servono can be taken as well as picked to act.

Learning Spark - Jules S. Damji 2020-07-16
Data is bigger, arrives faster, and comes in a variety of formats—and it all needs to be processed at scale for analytics or machine learning. But how can you process such varied workloads efficiently? Enter Apache Spark. Updated to include Spark 3.0, this second edition shows data engineers and data scientists why structure and unification in Spark matters. Specifically, this book explains how to perform simple and complex data analytics and employ machine learning algorithms. Through step-by-step walk-throughs, code snippets, and notebooks, you'll be able to: Learn Python, SQL, Scala, or Java high-level Structured APIs Understand Spark operations and SQL Engine Inspect, tune, and debug Spark operations with Spark configurations and Spark UI Connect to data sources: JSON, Parquet, CSV, Avro, ORC, Hive, S3, or Kafka Perform analytics on batch and streaming data using Structured Streaming Build reliable data pipelines with open source Delta Lake and Spark Develop machine learning pipelines with MLlib and productionize models using MLflow
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
IBM Platform Computing Solutions - Dino Quintero 2012-12-07
This IBM® Platform Computing Solutions Redbooks® publication is the first book to describe each of the available offerings that are part of the IBM portfolio of Cloud, analytics, and High Performance Computing (HPC) solutions for our clients. This IBM Redbooks publication delivers

descriptions of the available offerings from IBM Platform Computing that address challenges for our clients in each industry. We include a few implementation and testing scenarios with selected solutions. This publication helps strengthen the position of IBM Platform Computing solutions with a well-defined and documented deployment model within an IBM System x® environment. This deployment model offers clients a planned foundation for dynamic cloud infrastructure, provisioning, large-scale parallel HPC application development, cluster management, and grid applications. This IBM publication is targeted to IT specialists, IT architects, support personnel, and clients. This book is intended for anyone who wants information about how IBM Platform Computing solutions use IBM to provide a wide array of client solutions.

Handbook of Cloud Computing - Borko Furht 2010-09-11

Cloud computing has become a significant technology trend. Experts believe cloud computing is currently reshaping information technology and the IT marketplace. The advantages of using cloud computing include cost savings, speed to market, access to greater computing resources, high availability, and scalability. Handbook of Cloud Computing includes contributions from world experts in the field of cloud computing from academia, research laboratories and private industry. This book presents the systems, tools, and services of the leading providers of cloud computing; including Google, Yahoo, Amazon, IBM, and Microsoft. The basic concepts of cloud computing and cloud computing applications are also introduced. Current and future technologies applied in cloud computing are also discussed. Case studies, examples, and exercises are provided throughout. Handbook of Cloud Computing is intended for advanced-level students and researchers in computer science and electrical engineering as a reference book. This handbook is also beneficial to computer and system infrastructure designers, developers, business managers, entrepreneurs and investors within the cloud computing related industry.

Mastering Cloud Computing - Rajkumar Buyya 2013-04-05

Mastering Cloud Computing is designed for undergraduate students learning to develop cloud computing applications. Tomorrow's applications won't live on a single computer but will be deployed from and reside on a virtual server, accessible anywhere, any time. Tomorrow's application developers need to understand the requirements of building apps for these virtual systems, including concurrent programming, high-performance computing, and data-intensive systems. The book introduces the principles of distributed and parallel computing underlying cloud architectures and specifically focuses on virtualization, thread programming, task programming, and map-reduce programming. There are examples demonstrating all of these and more, with exercises and labs throughout. Explains how to make design choices and tradeoffs to consider when building applications to run in a virtual cloud environment Real-world case studies include scientific, business, and energy-efficiency considerations

Beginning Apache Pig - Balaswamy Vaddeman 2016-12-10

Learn to use Apache Pig to develop lightweight big data applications easily and quickly. This book shows you many optimization techniques and covers every context where Pig is used in big data analytics. Beginning Apache Pig shows you how Pig is easy to learn and requires relatively little time to develop big data applications.The book is divided into four parts: the complete features of Apache Pig; integration with other tools; how to solve complex business problems; and optimization of tools.You'll discover topics such as MapReduce and why it cannot meet every business need; the features of Pig Latin such as data types for each load, store, joins, groups, and ordering; how Pig workflows can be created; submitting Pig jobs using Hue; and working with Oozie. You'll also see how to extend the framework by writing UDFs and custom load, store, and filter functions. Finally you'll cover different optimization techniques such as gathering statistics about a Pig script, joining strategies, parallelism, and the role of data formats in good performance. What You Will Learn• Use all the features of Apache Pig• Integrate Apache Pig with other tools• Extend Apache Pig• Optimize Pig Latin code• Solve different use cases for Pig LatinWho This Book Is ForAll levels of IT professionals: architects, big data enthusiasts, engineers, developers, and big data administrators

Big Data con Hadoop - Gabriele Modena 2015-05-26T00:00:00+02:00
Hadoop è un progetto open source che permette di analizzare enormi quantità di dati distribuiti su cluster e file system differenti. Progettato per essere scalabile da un singolo server fino a migliaia di macchine, Hadoop si occupa anche di gestire problemi e guasti a livello applicativo - piuttosto che hardware - contribuendo a ottimizzare il mantenimento dei dati archiviati. Questo libro è dedicato a chi vuole entrare nel mondo della gestione e dell'analisi di Big Data. Attraverso l'uso degli strumenti e dei framework che compongo Hadoop 2, il lettore viene guidato nella progettazione e nell'implementazione di soluzioni di complessità differente, in grado di adattarsi a necessità operative e gestionali diverse che considerano sia la creazione e il mantenimento di dataset, sia la loro elaborazione e analisi per ottenere il massimo dai dati collezionati.
Introducing Microsoft SQL Server 2016 - Stacia Varga 2016-06-28
With Microsoft SQL Server 2016, a variety of new features and enhancements to the data platform deliver breakthrough performance, advanced security, and richer, integrated reporting and analytics capabilities. In this ebook, we introduce new security features: Always Encrypted, Row-Level Security, and dynamic data masking; discuss enhancements that enable you to better manage performance and storage: TemDB configuration, query store, and Stretch Database; review several improvements to Reporting Services; and also describe AlwaysOn Availability Groups, tabular enhancements, and R integration.
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.
Hadoop in Action - Chuck Lam 2010-11-30
Hadoop in Action teaches readers how to use Hadoop and write MapReduce programs. The intended readers are programmers, architects, and project managers who have to process large amounts of data offline. Hadoop in Action will lead the reader from obtaining a copy of Hadoop to setting it up in a cluster and writing data analytic programs. The book begins by making the basic idea of Hadoop and MapReduce easier to grasp by applying the default Hadoop installation to a few easy-to-follow tasks, such as analyzing changes in word frequency across a body of documents. The book continues through the basic concepts of MapReduce applications developed using Hadoop, including a close look at framework components, use of Hadoop for a variety of data analysis tasks, and numerous examples of Hadoop in action. Hadoop in Action will explain how to use Hadoop and present design patterns and practices of programming MapReduce. MapReduce is a complex idea both conceptually and in its implementation, and Hadoop users are challenged to learn all the knobs and levers for running Hadoop. This book takes you beyond the mechanics of running Hadoop, teaching you to write meaningful programs in a MapReduce framework. This book assumes the reader will have a basic familiarity with Java, as most code examples will be written in Java. Familiarity with basic statistical concepts (e.g. histogram, correlation) will help the reader appreciate the more advanced data processing examples.

Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book.

Big Data Now 2012 - O'Reilly Media Inc. 2014-08-14

The Big Data Now anthology is relevant to anyone who creates, collectsor relies upon data. It's not just a technical book or just a businessguide. Data is ubiquitous and it doesn't pay much attention toborders, so we've calibrated our coverage to follow it wherever itgoes.In the first edition of Big Data Now, the O'Reilly team tracked thebirth and early development of data tools and data science. Now, withthis second edition, we're seeing what happens when big data grows up:how it's being applied, where it's playing a role, and theconsequences -- good and bad alike -- of data's ascendance.We've organized the second edition of Big Data Now into five areas:Getting Up to Speed With Big Data -- Essential information on thestructures and definitions of big data.Big Data Tools, Techniques, and Strategies -- Expert guidance forturning big data theories into big data products.The Application of Big Data -- Examples of big data in action,including a look at the downside of data.What to Watch for in Big Data -- Thoughts on how big data will evolveand the role it will play across industries and domains.Big Data and Health Care -- A special section exploring thepossibilities that arise when data and health care come together.

Data Science and Analytics - Usha Batra 2020-05-27

This two-volume set (CCIS 1229 and CCIS 1230) constitutes the refereed proceedings of the 5th International Conference on Recent Developments in Science, Engineering and Technology, REDSET 2019, held in Gurugram, India, in November 2019. The 74 revised full papers presented were carefully reviewed and selected from total 353 submissions. The papers are organized in topical sections on data centric programming; next generation computing; social and web analytics; security in data science analytics; big data analytics.

Big data: cosa sono, come analizzarli e utilizzarli per fare marketing - Elisa Iandiorio 2021-03-15T00:00:00+01:00

«I big data sono come il sesso per gli adolescenti: tutti ne parlano, nessuno sa veramente come si fa, ma tutti pensano che gli altri lo fanno e allora dicono di farlo». Così scriveva Dan Ariely in un suo tweet del 2013. Oggi questa affermazione è ancora valida: il mondo dei big data interessa moltissimi aspetti della vita di un’azienda, ma non è ancora chiaro come approcciarsi a esso. Con questo libro ti invito a entrare nella post-adolescenza, acquisendo le conoscenze di base sui big data: cosa sono, come vengono utilizzati per aumentare le performance aziendali, come è possibile sviluppare una strategia attraverso l’analisi dei comportamenti d’acquisto del consumatore e quali cambiamenti il nuovo GDPR ha introdotto nel trattamento dei dati. L’obiettivo è aiutarti a capire, anche attraverso esempi concreti di aziende con cui ho affrontato il cammino dei big data, quali sono le fonti di dati più idonee per il tuo business e come utilizzarle per definire le tue buyer personas.

Hadoop 2 Quick-Start Guide - Douglas Eadline 2015-10-28

Get Started Fast with Apache Hadoop® 2, YARN, and Today’s Hadoop Ecosystem With Hadoop 2.x and YARN, Hadoop moves beyond MapReduce to become practical for virtually any type of data processing. Hadoop 2.x and the Data Lake concept represent a radical shift away from conventional approaches to data usage and storage. Hadoop 2.x installations offer unmatched scalability and breakthrough extensibility that supports new and existing Big Data analytics processing methods and models. Hadoop® 2 Quick-Start Guide is the first easy, accessible guide to Apache Hadoop 2.x, YARN, and the modern Hadoop ecosystem. Building on his unsurpassed experience teaching Hadoop and Big Data, author Douglas Eadline covers all the basics you need to know to install and use Hadoop 2 on personal computers or servers, and to navigate the powerful technologies that complement it. Eadline concisely introduces and explains every key Hadoop 2 concept, tool, and service, illustrating each with a simple “beginning-to-end” example and identifying trustworthy, up-to-date resources for learning more. This guide is ideal if you want to learn about Hadoop 2 without getting mired in technical details. Douglas Eadline will bring you up to speed quickly, whether you’re a user, admin, devops specialist, programmer, architect, analyst, or data scientist. Coverage Includes Understanding what Hadoop 2 and YARN do, and how they improve on Hadoop 1 with MapReduce

Understanding Hadoop-based Data Lakes versus RDBMS Data Warehouses Installing Hadoop 2 and core services on Linux machines, virtualized sandboxes, or clusters Exploring the Hadoop Distributed File System (HDFS) Understanding the essentials of MapReduce and YARN application programming Simplifying programming and data movement with Apache Pig, Hive, Sqoop, Flume, Oozie, and HBase Observing application progress, controlling jobs, and managing workflows Managing Hadoop efficiently with Apache Ambari–including recipes for HDFS to NFSv3 gateway, HDFS snapshots, and YARN configuration Learning basic Hadoop 2 troubleshooting, and installing Apache Hue and Apache Spark

Programming Hive - Edward Capriolo 2012-09-26

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

Data Analytics and Management - Ashish Khanna 2021-01-04

This book includes original unpublished contributions presented at the International Conference on Data Analytics and Management (ICDAM 2020), held at Jan Wyzykowski University, Poland, during June 2020. The book covers the topics in data analytics, data management, big data, computational intelligence, and communication networks. The book presents innovative work by leading academics, researchers, and experts from industry which is useful for young researchers and students.

Big Data Optimization: Recent Developments and Challenges - Ali Emrouznejad 2016-05-26

The main objective of this book is to provide the necessary background to work with big data by introducing some novel optimization algorithms and codes capable of working in the big data setting as well as introducing some applications in big data optimization for both academics and practitioners interested, and to benefit society, industry, academia, and government. Presenting applications in a variety of industries, this book will be useful for the researchers aiming to analyses large scale data. Several optimization algorithms for big data including convergent parallel algorithms, limited memory bundle algorithm, diagonal bundle method, convergent parallel algorithms, network analytics, and many more have been explored in this book.

Big data @l lavoro. Sfatare i miti, scoprire le opportunità - Davenport 2015

Retailing in the 21st Century - Manfred Krafft 2009-12-17

With crisp and insightful contributions from 47 of the world's leading experts in various facets of retailing, Retailing in the 21st Century offers in one book a compendium of state-of-the-art, cutting-edge knowledge to guide successful retailing in the new millennium. In our competitive world, retailing is an exciting, complex and critical sector of business in most developed as well as emerging economies. Today, the retailing industry is being buffeted by a number of forces simultaneously, for example the growth of online retailing and the advent of 'radio frequency identification' (RFID) technology. Making sense of it all is not easy but of vital importance to retailing practitioners, analysts and policymakers.

Cloud Computing - Dan C. Marinescu 2013-05-30

Cloud Computing: Theory and Practice provides students and IT professionals with an in-depth analysis of the cloud from the ground up. Beginning with a discussion of parallel computing and architectures and distributed systems, the book turns to contemporary cloud infrastructures, how they are being deployed at leading companies such as Amazon, Google and Apple, and how they can be applied in fields such as healthcare, banking and science. The volume also examines how to successfully deploy a cloud application across the enterprise using virtualization, resource management and the right amount of networking support, including content delivery networks and storage area networks. Developers will find a complete introduction to application development provided on a variety of platforms. Learn about recent trends in cloud computing in critical areas such as: resource management, security, energy consumption, ethics, and complex systems Get a detailed hands-on set of practical recipes that help simplify the deployment of a cloud based system for practical use of computing clouds along with an in-depth discussion of several projects Understand the evolution of cloud computing and why the cloud computing paradigm has a better chance

to succeed than previous efforts in large-scale distributed computing
Data Science and Big Data Analytics - EMC Education Services 2015-01-05
Data Science and Big Data Analytics is about harnessing the power of data for new insights. The book covers the breadth of activities and methods and tools that Data Scientists use. The content focuses on concepts, principles and practical applications that are applicable to any industry and technology environment, and the learning is supported and explained with examples that you can replicate using open-source software. This book will help you: Become a contributor on a data science team Deploy a structured lifecycle approach to data analytics problems Apply appropriate analytic techniques and tools to analyzing big data Learn how to tell a compelling story with data to drive business action Prepare for EMC Proven Professional Data Science Certification Corresponding data sets are available from the book's page at Wiley which you can find on the Wiley site by searching for the ISBN 9781118876138. Get started discovering, analyzing, visualizing, and presenting data in a meaningful way today!
Hadoop: The Definitive Guide - Tom White 2012-05-10
Ready to unlock the power of your data? With this comprehensive guide, you'll learn how to build and maintain reliable, scalable, distributed systems with Apache Hadoop. This book is ideal for programmers looking to analyze datasets of any size, and for administrators who want to set up and run Hadoop clusters. You'll find illuminating case studies that demonstrate how Hadoop is used to solve specific problems. This third edition covers recent changes to Hadoop, including material on the new MapReduce API, as well as MapReduce 2 and its more flexible execution model (YARN). Store large datasets with the Hadoop Distributed File System (HDFS) Run distributed computations with MapReduce Use Hadoop's data and I/O building blocks for compression, data integrity, serialization (including Avro), and persistence Discover common pitfalls and advanced features for writing real-world MapReduce programs Design, build, and administer a dedicated Hadoop cluster—or run Hadoop in the cloud Load data from relational databases into HDFS, using Sqoop Perform large-scale data processing with the Pig query language Analyze datasets with Hive, Hadoop's data warehousing system Take advantage of HBase for structured and semi-structured data, and ZooKeeper for building distributed systems
Real-Time Analytics - Byron Ellis 2014-06-23
Construct a robust end-to-end solution for analyzing and visualizing streaming data Real-time analytics is the hottest topic in data analytics today. In Real-Time Analytics: Techniques to Analyze and Visualize Streaming Data, expert Byron Ellis teaches data analysts technologies to build an effective real-time analytics platform. This platform can then be used to make sense of the constantly changing data that is beginning to outpace traditional batch-based analysis platforms. The author is among a very few leading experts in the field. He has a prestigious background in research, development, analytics, real-time visualization, and Big Data streaming and is uniquely qualified to help you explore this revolutionary field. Moving from a description of the overall analytic architecture of real-time analytics to using specific tools to obtain targeted results, Real-Time Analytics leverages open source and modern commercial tools to construct robust, efficient systems that can provide real-time analysis in a cost-effective manner. The book includes: A deep discussion of streaming data systems and architectures Instructions for analyzing, storing, and delivering streaming data Tips on aggregating data and working with sets Information on data warehousing options and techniques Real-Time Analytics includes in-depth case studies for website analytics, Big Data, visualizing streaming and mobile data, and mining and visualizing operational data flows. The book's "recipe" layout lets readers quickly learn and implement different techniques. All of the code examples presented in the book, along with their related data sets, are available on the companion website.
AI and Big Data on IBM Power Systems Servers - Scott Vetter 2019-04-10
As big data becomes more ubiquitous, businesses are wondering how they can best leverage it to gain insight into their most important business questions. Using machine learning (ML) and deep learning (DL)

in big data environments can identify historical patterns and build artificial intelligence (AI) models that can help businesses to improve customer experience, add services and offerings, identify new revenue streams or lines of business (LOBs), and optimize business or manufacturing operations. The power of AI for predictive analytics is being harnessed across all industries, so it is important that businesses familiarize themselves with all of the tools and techniques that are available for integration with their data lake environments. In this IBM® Redbooks® publication, we cover the best practices for deploying and integrating some of the best AI solutions on the market, including: IBM Watson Machine Learning Accelerator (see note for product naming) IBM Watson Studio Local IBM Power SystemsTM IBM SpectrumTM Scale IBM Data Science Experience (IBM DSX) IBM Elastic StorageTM Server Hortonworks Data Platform (HDP) Hortonworks DataFlow (HDF) H2O Driverless AI We map out all the integrations that are possible with our different AI solutions and how they can integrate with your existing or new data lake. We also walk you through some of our client use cases and show you how some of the industry leaders are using Hortonworks, IBM PowerAI, and IBM Watson Studio Local to drive decision making. We also advise you on your deployment options, when to use a GPU, and why you should use the IBM Elastic Storage Server (IBM ESS) to improve storage management. Lastly, we describe how to integrate IBM Watson Machine Learning Accelerator and Hortonworks with or without IBM Watson Studio Local, how to access real-time data, and security. Note: IBM Watson Machine Learning Accelerator is the new product name for IBM PowerAI Enterprise. Note: Hortonworks merged with Cloudera in January 2019. The new company is called Cloudera. References to Hortonworks as a business entity in this publication are now referring to the merged company. Product names beginning with Hortonworks continue to be marketed and sold under their original names.

Mining of Massive Datasets - Jure Leskovec 2014-11-13

Now in its second edition, this book focuses on practical algorithms for mining data from even the largest datasets.

Big Data - Alessandro Rezzani 2013-10-01

Ogni giorno nel mondo vengono creati miliardi di dati digitali. Questa mole di informazione proviene dal notevole incremento di dispositivi che automatizzano numerose operazioni – record delle transazioni di acquisto e segnali GPS dei cellulari, per esempio – e dal Web: foto, video, post, articoli e contenuti digitali generati e diffusi dagli utenti tramite i social media. L'elaborazione di questi "big data" richiede elevate capacità di calcolo, tecnologie e risorse che vanno ben al di là dei sistemi convenzionali di gestione e immagazzinamento dei dati. Il testo esplora il mondo dei "grandi dati" e ne offre una descrizione e classificazione, presentando le opportunità che possono derivare dal loro utilizzo. Descrive le soluzioni software e hardware dedicate, riservando ampio spazio alle implementazioni Open Source e alle principali offerte cloud. Si propone dunque come una guida approfondita agli strumenti e alle tecnologie che permettono l'analisi e la gestione di grandi quantità di dati. Il volume è dedicato a chi, in università e in azienda (database administrator, IT manager, professionisti di Business Intelligence) intende approfondire le tematiche relative ai big data. È, inoltre, un valido supporto per il management aziendale per comprendere come ottenere informazioni utilizzabili nei processi decisionali. Alessandro Rezzani insegna presso l'Università Bocconi di Milano. È esperto di progettazione e implementazione di Data Warehouse, di processi ETL, database multidimensionali e soluzioni di reporting. Attualmente si occupa di disegno e implementazione di soluzioni di Business Intelligence presso Factory Software. Con Apogeo Education ha pubblicato "Business Intelligence. Processi, metodi, utilizzo in azienda", 2012.

HBase - Lars George 2011-09-05

"HBase: The Definitive Guide" provides the details for evaluating this high-performance, non-relational database, or putting it into practice right away. HBase's adoption rate is beginning to climb, and IT executives are asking pointed questions about this high-capacity database.

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

Big Data Systems - Jawwad Ahmed Shamsi 2021-05-10

Big Data Systems encompass massive challenges related to data diversity, storage mechanisms, and requirements of massive computational power. Further, capabilities of big data systems also vary with respect to type of problems. For instance, distributed memory systems are not recommended for iterative algorithms. Similarly, variations in big data systems also exist related to consistency and fault tolerance. The purpose of this book is to provide a detailed explanation of big data systems. The book covers various topics including Networking, Security, Privacy, Storage, Computation, Cloud Computing, NoSQL and NewSQL systems, High Performance Computing, and Deep Learning. An illustrative and practical approach has been adopted in which theoretical topics have been aided by well-explained programming and illustrative examples. Key Features: Introduces concepts and evolution of Big Data technology. Illustrates examples for thorough understanding. Contains programming examples for hands on development. Explains a variety of topics including NoSQL Systems, NewSQL systems, Security, Privacy, Networking, Cloud, High Performance Computing, and Deep Learning. Exemplifies widely used big data technologies such as Hadoop and Spark. Includes discussion on case studies and open issues. Provides end of chapter questions for enhanced learning.

Practical Statistics for Data Scientists - Peter Bruce 2017-05-10

Statistical methods are a key part of of data science, yet very few data scientists have any formal statistics training. Courses and books on basic statistics rarely cover the topic from a data science perspective. This practical guide explains how to apply various statistical methods to data science, tells you how to avoid their misuse, and gives you advice on what's important and what's not. Many data science resources incorporate statistical methods but lack a deeper statistical perspective. If you’re familiar with the R programming language, and have some exposure to statistics, this quick reference bridges the gap in an accessible, readable format. With this book, you’ll learn: Why exploratory data analysis is a key preliminary step in data science How random

sampling can reduce bias and yield a higher quality dataset, even with big data How the principles of experimental design yield definitive answers to questions How to use regression to estimate outcomes and detect anomalies Key classification techniques for predicting which categories a record belongs to Statistical machine learning methods that “learn” from data Unsupervised learning methods for extracting meaning from unlabeled data

Python for R Users - Ajay Ohri 2017-11-13

The definitive guide for statisticians and data scientists who understand the advantages of becoming proficient in both R and Python The first book of its kind, Python for R Users: A Data Science Approach makes it easy for R programmers to code in Python and Python users to program in R. Short on theory and long on actionable analytics, it provides readers with a detailed comparative introduction and overview of both languages and features concise tutorials with command-by-command translations—complete with sample code—of R to Python and Python to R. Following an introduction to both languages, the author cuts to the chase with step-by-step coverage of the full range of pertinent programming features and functions, including data input, data inspection/data quality, data analysis, and data visualization. Statistical modeling, machine learning, and data mining—including supervised and unsupervised data mining methods—are treated in detail, as are time series forecasting, text mining, and natural language processing. • Features a quick-learning format with concise tutorials and actionable analytics • Provides command-by-command translations of R to Python and vice versa • Incorporates Python and R code throughout to make it easier for readers to compare and contrast features in both languages • Offers numerous comparative examples and applications in both programming languages • Designed for use for practitioners and students that know one language and want to learn the other • Supplies slides useful for teaching and learning either software on a companion website Python for R Users: A Data Science Approach is a valuable working resource for computer scientists and data scientists that know R and would like to learn Python or are familiar with Python and want to learn R. It also functions as textbook for students of computer science and statistics. A. Ohri is the founder of Decisionstats.com and currently works as a senior data scientist. He has advised multiple startups in analytics off-shoring, analytics services, and analytics education, as well as using social media to enhance buzz for analytics products. Mr. Ohri's research interests include spreading open source analytics, analyzing social media manipulation with mechanism design, simpler interfaces for cloud computing, investigating climate change and knowledge flows. His other books include R for Business Analytics and R for Cloud Computing.

Mathematics for Machine Learning - Marc Peter Deisenroth 2020-04-23

The fundamental mathematical tools needed to understand machine learning include linear algebra, analytic geometry, matrix decompositions, vector calculus, optimization, probability and statistics. These topics are traditionally taught in disparate courses, making it hard for data science or computer science students, or professionals, to efficiently learn the mathematics. This self-contained textbook bridges the gap between mathematical and machine learning texts, introducing the mathematical concepts with a minimum of prerequisites. It uses these concepts to derive four central machine learning methods: linear regression, principal component analysis, Gaussian mixture models and support vector machines. For students and others with a mathematical background, these derivations provide a starting point to machine learning texts. For those learning the mathematics for the first time, the methods help build intuition and practical experience with applying mathematical concepts. Every chapter includes worked examples and exercises to test understanding. Programming tutorials are offered on the book's web site.

Learning PySpark - Tomasz Drabas 2017-02-27

Build data-intensive applications locally and deploy at scale using the combined powers of Python and Spark 2.0 About This Book Learn why and how you can efficiently use Python to process data and build machine learning models in Apache Spark 2.0 Develop and deploy efficient, scalable real-time Spark solutions Take your understanding of

using Spark with Python to the next level with this jump start guide Who This Book Is For If you are a Python developer who wants to learn about the Apache Spark 2.0 ecosystem, this book is for you. A firm understanding of Python is expected to get the best out of the book. Familiarity with Spark would be useful, but is not mandatory. What You Will Learn Learn about Apache Spark and the Spark 2.0 architecture Build and interact with Spark DataFrames using Spark SQL Learn how to solve graph and deep learning problems using GraphFrames and TensorFrames respectively Read, transform, and understand data and use it to train machine learning models Build machine learning models with MLlib and ML Learn how to submit your applications programmatically using spark-submit Deploy locally built applications to a cluster In Detail Apache Spark is an open source framework for efficient cluster computing with a strong interface for data parallelism and fault tolerance. This book will show you how to leverage the power of Python and put it to use in the Spark ecosystem. You will start by getting a firm understanding of the Spark 2.0 architecture and how to set up a Python environment for Spark. You will get familiar with the modules available in PySpark. You will learn how to abstract data with RDDs and DataFrames and understand the streaming capabilities of PySpark. Also, you will get a thorough overview of machine learning capabilities of PySpark using ML and MLlib, graph processing using GraphFrames, and polyglot persistence using Blaze. Finally, you will learn how to deploy your applications to the cloud using the spark-submit command. By the end of this book, you will have established a firm understanding of the Spark Python API and how it can be used to build data-intensive applications. Style and approach This book takes a very comprehensive, step-by-step approach so you understand how the Spark ecosystem can be used with Python to develop efficient, scalable solutions. Every chapter is standalone and written in a very easy-to-understand manner, with a focus on both the hows and the whys of each concept.

Big Data Security - Shibakali Gupta 2019-10-08

THE SERIES: FRONTIERS IN COMPUTATIONAL INTELLIGENCE The series Frontiers In Computational Intelligence is envisioned to provide comprehensive coverage and understanding of cutting edge research in computational intelligence. It intends to augment the scholarly discourse on all topics relating to the advances in artifi cial life and machine learning in the form of metaheuristics, approximate reasoning, and robotics. Latest research findings are coupled with applications to varied domains of engineering and computer sciences. This field is steadily growing especially with the advent of novel machine learning algorithms being applied to different domains of engineering and technology. The series brings together leading researchers that intend to continue to advance the field and create a broad knowledge about the most recent research. Series Editor Dr. Siddhartha Bhattacharyya, CHRIST (Deemed to be University), Bangalore, India Editorial Advisory Board Dr. Elizabeth Behrman, Wichita State University, Kansas, USA Dr. Goran Klepac Dr. Leo Mrsic, Algebra University College, Croatia Dr. Aboul Ella Hassanien, Cairo University, Egypt Dr. Jan Platos, VSB-Technical University of Ostrava, Czech Republic Dr. Xiao-Zhi Gao, University of Eastern Finland, Finland Dr. Wellington Pinheiro dos Santos, Federal University of Pernambuco, Brazil

Network Security Through Data Analysis - Michael Collins 2014-02-10

In this practical guide, security researcher Michael Collins shows you several techniques and tools for collecting and analyzing network traffic datasets. You'll understand how your network is used, and what actions are necessary to protect and improve it. Divided into three sections, this book examines the process of collecting and organizing data, various tools for analysis, and several different analytic scenarios and techniques.

Trino: The Definitive Guide - Matt Fuller 2021-04-14

Perform fast interactive analytics against different data sources using the Trino high-performance distributed SQL query engine. With this practical guide, you'll learn how to conduct analytics on data where it lives, whether it's Hive, Cassandra, a relational database, or a proprietary data store. Analysts, software engineers, and production engineers will learn how to manage, use, and even develop with Trino. Initially developed by

Facebook, open source Trino is now used by Netflix, Airbnb, LinkedIn, Twitter, Uber, and many other companies. Matt Fuller, Manfred Moser, and Martin Traverso show you how a single Trino query can combine data from multiple sources to allow for analytics across your entire organization. Get started: Explore Trino's use cases and learn about tools that will help you connect to Trino and query data Go deeper: Learn Trino's internal workings, including how to connect to and query data sources with support for SQL statements, operators, functions, and more Put Trino in production: Secure Trino, monitor workloads, tune queries, and connect more applications; learn how other organizations apply Trino

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

Fintech with Artificial Intelligence, Big Data, and Blockchain - Paul Moon Sub Choi 2021-03-08

This book introduces readers to recent advancements in financial technologies. The contents cover some of the state-of-the-art fields in financial technology, practice, and research associated with artificial intelligence, big data, and blockchain—all of which are transforming the nature of how products and services are designed and delivered, making less adaptable institutions fast become obsolete. The book provides the fundamental framework, research insights, and empirical evidence in the efficacy of these new technologies, employing practical and academic approaches to help professionals and academics reach innovative solutions and grow competitive strengths.

Big Data and Business Analytics - Jay Liebowitz 2016-04-19

"The chapters in this volume offer useful case studies, technical roadmaps, lessons learned, and a few prescriptions todo this, avoid that.'"-From the Foreword by Joe LaCugna, Ph.D., Enterprise Analytics and Business Intelligence, Starbucks Coffee CompanyWith the growing barrage of "big data," it becomes vitally important for organizations to mak

Programming Pig - Alan Gates 2011-10-06

This guide is an ideal learning tool and reference for Apache Pig, the programming language that helps programmers describe and run large data projects on Hadoop. With Pig, they can analyze data without having to create a full-fledged application--making it easy for them to experiment with new data sets.