Big Data Now 2012 Edition Ebook Oreilly Media

This is likewise one of the factors by obtaining the soft documents of this **big data now 2012 edition ebook oreilly media** by online. You might not require more become old to spend to go to the books instigation as with ease as search for them. In some cases, you likewise get not discover the declaration big data now 2012 edition ebook oreilly media that you are looking for. It will definitely squander the time.

However below, behind you visit this web page, it will be as a result totally simple to acquire as capably as download lead big data now 2012 edition ebook oreilly media

It will not undertake many mature as we accustom before. You can realize it while show something else at home and even in your workplace. consequently easy! So, are you question? Just exercise just what we present below as well as review **big data now 2012 edition ebook oreilly media** what you as soon as to read!

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
Big Data Now: Current Perspectives from O'Reilly Radar - O'Reilly Radar Team 2011-08-30
This collection represents the full spectrum of data-related content we've published on O'Reilly Radar over the last year. Mike Loukides kicked things off in June 2010 with "What is data science?" and from there we've pursued the various threads and themes that naturally emerged. Now, roughly a year later, we can look back over all we've covered and identify a number of core data areas: Data issues -- The opportunities and ambiguities of the data space are evident in discussions around privacy, the implications of data-centric industries, and the debate about the phrase "data science" itself. The application of data: products and processes – A "data product" can emerge from virtually any domain, including everything from data startups to established enterprises to media/journalism to education and research. Data science and data tools -- The tools and technologies that drive data science are of course essential to this space, but the varied techniques being applied are also key to understanding the big data arena. The business of data – Take a closer look at the actions connected to data -- the finding, organizing, and analyzing that provide organizations of all sizes with the information they need to compete.
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
Big Data Now - O'Reilly (Firm) 2011
The business of data -- take a closer look at the actions connected to data -- the finding, organizing, and analyzing that provide organizations of all sizes with the information they need to compete.
Big Data - Viktor Mayer-Schönberger 2013
This revelatory exploration of big data, which refers to our newfound ability to crunch vast amounts of information, analyze it instantly and draw profound and surprising conclusions from it, discusses how it will change our lives and what we can do to protect ourselves from its hazards. 75,000 first printing.
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
Spring Data - Mark Pollack 2012-10-24
You can choose several data access frameworks when building Java enterprise applications that work with relational databases. But what about big data? This hands-on introduction shows you how Spring Data makes it relatively easy to build applications across a wide range of new data access technologies such as NoSQL and Hadoop. Through several sample projects, you'll learn how Spring Data provides a consistent programming model that retains NoSQL-specific features and capabilities, and helps you develop Hadoop applications across a wide range of use-cases such as data analysis, event stream processing, and workflow. You'll also discover the features Spring Data adds to Spring's existing JPA and JDBC support for writing RDBMS-based data access layers. Learn about Spring's template helper classes to simplify the use of database-specific functionality Explore Spring Data's repository abstraction and advanced query functionality Use Spring Data with Redis (key/value store), HBase (column-family), MongoDB (document database), and Neo4j (graph database) Discover the GemFire distributed data grid solution Export Spring Data JPA-managed entities to the Web as RESTful web services Simplify the development of HBase applications, using a lightweight object-mapping framework Build example big-data pipelines with Spring Batch and Spring Integration
Bad Data Handbook - Q. Ethan McCallum 2012-11-07
What is bad data? Some people consider it a technical phenomenon, like missing values or malformed records, but bad data includes a lot more. In this handbook, data expert Q. Ethan McCallum has gathered 19 colleagues from every corner of the data arena to reveal how they've recovered from nasty data problems. From cranky storage to poor representation to misguided policy, there are many paths to bad data. Bottom line? Bad data is data that gets in the way. This book explains effective ways to get around it. Among the many topics covered, you'll discover how to: Test drive your data to see if it's ready for analysis Work spreadsheet data into a usable form Handle encoding problems that lurk

Downloaded from hoekstratruck.com on by guest

in text data Develop a successful web-scraping effort Use NLP tools to reveal the real sentiment of online reviews Address cloud computing issues that can impact your analysis effort Avoid policies that create data analysis roadblocks Take a systematic approach to data quality analysis

Big Data Now: 2012 Edition - O'Reilly Media, Inc. 2012-10-23

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.

Version Control with Git - Jon Loeliger 2012-08-14

Get up to speed on Git for tracking, branching, merging, and managing code revisions. Through a series of step-by-step tutorials, this practical guide takes you quickly from Git fundamentals to advanced techniques, and provides friendly yet rigorous advice for navigating the many functions of this open source version control system. This thoroughly revised edition also includes tips for manipulating trees, extended coverage of the reflog and stash, and a complete introduction to the GitHub repository. Git lets you manage code development in a virtually endless variety of ways, once you understand how to harness the system’s flexibility. This book shows you how. Learn how to use Git for several real-world development scenarios Gain insight into Git’s common-use cases, initial tasks, and basic functions Use the system for both centralized and distributed version control Learn how to manage merges, conflicts, patches, and diffs Apply advanced techniques such as rebasing, hooks, and ways to handle submodules Interact with Subversion (SVN) repositories—including SVN to Git conversions Navigate, use, and contribute to open source projects though GitHub

Planning for Big Data - Edd Dumbill 2014-07-31

In an age where everything is measurable, understanding big data is an essential. From creating new data-driven products through to increasing operational efficiency, big data has the potential to make your organization both more competitive and more innovative.As this emerging field transitions from the bleeding edge to enterprise infrastructure, it's vital to understand not only the technologies involved, but the organizational and cultural demands of being data-driven.

Privacy and Big Data - Terence Craig 2011-09-23

"The players, regulators, and stakeholders"--Cover.

Getting Started with Storm - Jonathan Leibiusky 2012-08-31

Even as big data is turning the world upside down, the next phase of the revolution is already taking shape: real-time data analysis. This hands-on guide introduces you to Storm, a distributed, JVM-based system for processing streaming data. Through simple tutorials, sample Java code, and a complete real-world scenario, you’ll learn how to build fast, fault-tolerant solutions that process results as soon as the data arrives. Discover how easy it is to set up Storm clusters for solving various problems, including continuous data computation, distributed remote procedure calls, and data stream processing. Learn how to program Storm components: spouts for data input and bolts for data transformation Discover how data is exchanged between spouts and bolts in a Storm topology Make spouts fault-tolerant with several commonly used design strategies Explore bolts—their life cycle, strategies for design, and ways to implement them Scale your solution by defining each component’s level of parallelism Study a real-time web analytics system built with Node.js, a Redis server, and a Storm topology Write spouts and bolts with non-JVM languages such as Python, Ruby, and Javascript

Programming Hive - Edward Capriolo 2012-09-26

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

The Twitter Book - Tim O'Reilly 2011-11-07

Twitter is not just for talking about your breakfast anymore. It’s become an indispensable communications tool for businesses, non-profits, celebrities, and people around the globe. With the second edition of this friendly, full-color guide, you’ll quickly get up to speed not only on standard features, but also on new options and nuanced uses that will help you tweet with confidence. Co-written by two widely recognized Twitter experts, The Twitter Book is packed with all-new real-world examples, solid advice, and clear explanations guaranteed to turn you into a power user. Use Twitter to connect with colleagues, customers, family, and friends Stand out on Twitter Avoid common gaffes and pitfalls Build a critical communications channel with Twitter—and use the best third-party tools to manage it. Want to learn how to use Twitter like a pro? Get the book that readers and critics alike rave about.

Deep Learning - Josh Patterson 2017-07-28

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

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

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

Building Data Science Teams - DJ Patil 2011-09-15

As data science evolves to become a business necessity, the importance of assembling a strong and innovative data teams grows. In this in-depth report, data scientist DJ Patil explains the skills, perspectives, tools and processes that position data science teams for success. Topics include: What it means to be "data driven." The unique roles of data scientists. The four essential qualities of data scientists. Patil's first-hand experience building the LinkedIn data science team.

Lean Analytics - Alistair Croll 2013-04-15

Offers six sample business models and thirty case studies to help build and monetize a business.

Enterprise Search - Martin White 2015-10-13

Is your organization rapidly accumulating more information than you know how to manage? This updated edition of Enterprise Search helps you create an enterprise search solution based on more than just technology. Author Martin White shows you how to plan and implement a managed search environment that meets the needs of your business and your employees. You'll learn why it’s absolutely vital to have a dedicated staff manage your search technology and support your users. New material for this second edition includes material on SharePoint 2013 search, managing open source search development, website search, designing the search user, and assessing search performance. Chapters now include a Further Reading section for computer science and

Downloaded from hoekstratruck.com on by guest

information science students. Topics include: 10 critical success factors to assess organizational search maturity Essential skills needed to support a successful search application How to specify and manage open source search development How to manage SharePoint 2013 search Methods to assess the business impact of search Best practices in user interface design The importance of search for websites What to include in a search strategy

Programming Google App Engine - Dan Sanderson 2009-11-23

As one of today's cloud computing services, Google App Engine does more than provide access to a large system of servers. It also offers you a simple model for building applications that scale automatically to accommodate millions of users. With Programming Google App Engine, you'll get expert practical guidance that will help you make the best use of this powerful platform. Google engineer Dan Sanderson shows you how to design your applications for scalability, including ways to perform common development tasks using App Engine's APIs and scalable services. You'll learn about App Engine's application server architecture, runtime environments, and scalable datastore for distributing data, as well as techniques for optimizing your application. App Engine offers nearly unlimited computing power, and this book provides clear and concise instructions for getting the most from it right from the source. Discover the differences between traditional web development and development with App Engine Learn the details of App Engine's Python and Java runtime environments Understand how App Engine handles web requests and executes application code Learn how to use App Engine's scalable datastore, including queries and indexes, transactions, and data modeling Use task queues to parallelize and distribute work across the infrastructure Deploy and manage applications with ease

Statistics in a Nutshell - Sarah Boslaugh 2012-11-15

A clear and concise introduction and reference for anyone new to the subject of statistics.

R in a Nutshell - Joseph Adler 2012-09-26

If you're considering R for statistical computing and data visualization, this book provides a quick and practical guide to just about everything you can do with the open source R language and software environment. You'll learn how to write R functions and use R packages to help you prepare, visualize, and analyze data. Author Joseph Adler illustrates each process with a wealth of examples from medicine, business, and sports. Updated for R 2.14 and 2.15, this second edition includes new and expanded chapters on R performance, the ggplot2 data visualization package, and parallel R computing with Hadoop. Get started quickly with an R tutorial and hundreds of examples Explore R syntax, objects, and other language details Find thousands of user-contributed R packages online, including Bioconductor Learn how to use R to prepare data for analysis Visualize your data with R's graphics, lattice, and ggplot2 packages Use R to calculate statistical fests, fit models, and compute probability distributions Speed up intensive computations by writing parallel R programs for Hadoop Get a complete desktop reference to R

Big Data Now - O'Reilly Media 2012

Disruptive Possibilities: How Big Data Changes Everything - Jeffrey Needham 2013-05-06

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

PostgreSQL: Up and Running - Regina O. Obe 2012-07-06

Thinking of migrating to PostgreSQL? This updated guide helps you quickly understand and use the 9.3 release of this open source database system. You'll not only learn about its unique enterprise-class features, but also discover that PostgeSQL is more than just a database system—it's also an impressive application platform. Using numerous examples, this book shows you how to achieve tasks that are difficult or impossible in other databases. The second edition covers LATERAL queries, augmented JSON support, materialized views, and other key topics. If you're an existing PostgreSQL user, you'll pick up gems you may have missed along the way. Learn basic administration tasks, such as role management, database creation, backup, and restore Apply the psql command-line utility and the pgAdmin graphical administration tool Explore PostgreSQL tables, constraints, and indexes Learn powerful SQL constructs not generally found in other databases Use several different languages to write database functions Tune your queries to run as fast as your hardware will allow Query external and variegated data sources with Foreign Data Wrappers Learn how to replicate data, using built-in replication features

Designing Great Data Products - Jeremy Howard 2012-03-23

In the past few years, we've seen many data products based on predictive modeling. These products range from weather forecasting to recommendation engines like Amazon's. Prediction technology can be interesting and mathematically elegant, but we need to take the next step: going from recommendations to products that can produce optimal strategies for meeting concrete business objectives. We already know how to build these products: they've been in use for the past decade or so, but they're not as common as they should be. This report shows how to take the next step: to go from simple predictions and recommendations to a new generation of data products with the potential to revolutionize entire industries.

What Is Data Science? - Mike Loukides 2011-04-10

We've all heard it: according to Hal Varian, statistics is the next sexy job. Five years ago, in What is Web 2.0, Tim O'Reilly said that "data is the next Intel Inside." But what does that statement mean? Why do we suddenly care about statistics and about data? This report examines the many sides of data science -- the technologies, the companies and the unique skill sets.The web is full of "data-driven apps." Almost any e-commerce application is a data-driven application. There's a database behind a web front end, and middleware that talks to a number of other databases and data services (credit card processing companies, banks, and so on). But merely using data isn't really what we mean by "data science." A data application acquires its value from the data itself, and creates more data as a result. It's not just an application with data; it's a data product. Data science enables the creation of data products.

97 Things Every Cloud Engineer Should Know - Emily Freeman 2020-12-04

If you create, manage, operate, or configure systems running in the cloud, you're a cloud engineer--even if you work as a system administrator, software developer, data scientist, or site reliability engineer. With this book, professionals from around the world provide valuable insight into today's cloud engineering role. These concise articles explore the entire cloud computing experience, including fundamentals, architecture, and migration. You'll delve into security and compliance, operations and reliability, and software development. And examine networking, organizational culture, and more. You're sure to find 1, 2, or 97 things that inspire you to dig deeper and expand your own career. "Three Keys to Making the Right Multicloud Decisions," Brendan O'Leary "Serverless Bad Practices," Manases Jesus Galindo Bello "Failing a Cloud Migration," Lee Atchison "Treat Your Cloud Environment as If It Were On Premises," Iyana Garry "What Is Toil, and Why Are SREs Obsessed with It?", Zachary Nickens "Lean QA: The QA Evolving in the DevOps World," Theresa Neate "How Economies of Scale Work in the Cloud," Jon Moore "The Cloud Is Not About the Cloud," Ken Corless "Data Gravity: The Importance of Data Management in the Cloud," Geoff Hughes "Even in the Cloud, the Network Is the Foundation," David Murray "Cloud Engineering Is About Culture, Not Containers," Holly Cummins

Streaming Systems - Tyler Akidau 2018-07-16

Streaming data is a big deal in big data these days. As more and more businesses seek to tame the massive unbounded data sets that pervade our world, streaming systems have finally reached a level of maturity sufficient for mainstream adoption. With this practical guide, data engineers, data scientists, and developers will learn how to work with streaming data in a conceptual and platform-agnostic way. Expanded from Tyler Akidau's popular blog posts "Streaming 101" and "Streaming 102", this book takes you from an introductory level to a nuanced understanding of the what, where, when, and how of processing real-time data streams. You'll also dive deep into watermarks and exactly-once processing with co-authors Slava Chernyak and Reuven Lax. You'll explore: How streaming and batch data processing patterns compare The core principles and concepts behind robust out-of-order data processing How watermarks track progress and completeness in infinite datasets

Downloaded from hoekstratruck.com on by guest

How exactly-once data processing techniques ensure correctness How the concepts of streams and tables form the foundations of both batch and streaming data processing The practical motivations behind a powerful persistent state mechanism, driven by a real-world example How time-varying relations provide a link between stream processing and the world of SQL and relational algebra

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

Python for Data Analysis - Wes McKinney 2017-09-25

Get complete instructions for manipulating, processing, cleaning, and crunching datasets in Python. Updated for Python 3.6, the second edition of this hands-on guide is packed with practical case studies that show you how to solve a broad set of data analysis problems effectively. You'll learn the latest versions of pandas, NumPy, IPython, and Jupyter in the process. Written by Wes McKinney, the creator of the Python pandas project, this book is a practical, modern introduction to data science tools in Python. It's ideal for analysts new to Python and for Python programmers new to data science and scientific computing. Data files and related material are available on GitHub. Use the IPython shell and Jupyter notebook for exploratory computing Learn basic and advanced features in NumPy (Numerical Python) Get started with data analysis tools in the pandas library Use flexible tools to load, clean, transform, merge, and reshape data Create informative visualizations with matplotlib Apply the pandas groupby facility to slice, dice, and summarize datasets Analyze and manipulate regular and irregular time series data Learn how to solve real-world data analysis problems with thorough, detailed examples

Natural Language Annotation for Machine Learning - James Pustejovsky 2012-10-11

Create your own natural language training corpus for machine learning. Whether you're working with English, Chinese, or any other natural language, this hands-on book guides you through a proven annotation development cycle—the process of adding metadata to your training corpus to help ML algorithms work more efficiently. You don't need any programming or linguistics experience to get started. Using detailed examples at every step, you'll learn how the MATTER Annotation Development Process helps you Model, Annotate, Train, Test, Evaluate, and Revise your training corpus. You also get a complete walkthrough of a real-world annotation project. Define a clear annotation goal before collecting your dataset (corpus) Learn tools for analyzing the linguistic content of your corpus Build a model and specification for your annotation project Examine the different annotation formats, from basic XML to the Linguistic Annotation Framework Create a gold standard corpus that can be used to train and test ML algorithms Select the ML algorithms that will process your annotated data Evaluate the test results and revise your annotation task Learn how to use lightweight software for annotating texts and adjudicating the annotations This book is a perfect companion to O'Reilly's Natural Language Processing with Python.

Spark: The Definitive Guide - Bill Chambers 2018-02-08

Learn how to use, deploy, and maintain Apache Spark with this comprehensive guide, written by the creators of the open-source cluster-computing framework. With an emphasis on improvements and new features in Spark 2.0, authors Bill Chambers and Matei Zaharia break down Spark topics into distinct sections, each with unique goals. You'll explore the basic operations and common functions of Spark's structured APIs, as well as Structured Streaming, a new high-level API for building end-to-end streaming applications. Developers and system administrators will learn the fundamentals of monitoring, tuning, and debugging Spark, and explore machine learning techniques and scenarios for employing MLlib, Spark's scalable machine-learning library. Get a gentle overview of big data and Spark Learn about DataFrames, SQL, and Datasets—Spark's core APIs—through worked examples Dive into Spark's low-level APIs, RDDs, and execution of SQL and DataFrames Understand how Spark runs on a cluster Debug, monitor, and tune Spark clusters and applications Learn the power of Structured Streaming, Spark's stream-processing engine Learn how you can apply MLlib to a variety of problems, including classification or recommendation

Python Cookbook - David Beazley 2013-05-10

If you need help writing programs in Python 3, or want to update older Python 2 code, this book is just the ticket. Packed with practical recipes written and tested with Python 3.3, this unique cookbook is for experienced Python programmers who want to focus on modern tools and idioms. Inside, youâ??ll find complete recipes for more than a dozen topics, covering the core Python language as well as tasks common to a wide variety of application domains. Each recipe contains code samples you can use in your projects right away, along with a discussion about how and why the solution works. Topics include: Data Structures and Algorithms Strings and Text Numbers, Dates, and Times Iterators and Generators Files and I/O Data Encoding and Processing Functions Classes and Objects Metaprogramming Modules and Packages Network and Web Programming Concurrency Utility Scripting and System Administration Testing, Debugging, and Exceptions C Extensions

Python and HDF5 - Andrew Collette 2013-10-21

Gain hands-on experience with HDF5 for storing scientific data in Python. This practical guide quickly gets you up to speed on the details, best practices, and pitfalls of using HDF5 to archive and share numerical datasets ranging in size from gigabytes to terabytes. Through real-world examples and practical exercises, you'll explore topics such as scientific datasets, hierarchically organized groups, user-defined metadata, and interoperable files. Examples are applicable for users of both Python 2 and Python 3. If you're familiar with the basics of Python data analysis, this is an ideal introduction to HDF5. Get set up with HDF5 tools and create your first HDF5 file Work with datasets by learning the HDF5 Dataset object Understand advanced features like dataset chunking and compression Learn how to work with HDF5's hierarchical structure, using groups Create self-describing files by adding metadata with HDF5 attributes Take advantage of HDF5's type system to create interoperable files Express relationships among data with references, named types, and dimension scales Discover how Python mechanisms for writing parallel code interact with HDF5

Real-Time Big Data Analytics: Emerging Architecture - Mike Barlow 2013-06-24

Five or six years ago, analysts working with big datasets made queries and got the results back overnight. The data world was revolutionized a few years ago when Hadoop and other tools made it possible to getthe results from queries in minutes. But the revolution continues. Analysts now demand sub-second, near real-time query results. Fortunately, we have the tools to deliver them. This report examines tools and technologies that are driving real-time big data analytics.

Enterprise Analytics - Thomas H. Davenport 2013

"International Institute for Analytics"--Dust jacket.

Learning R - Richard Cotton 2013-09-09

Learn how to perform data analysis with the R language and software environment, even if you have little or no programming experience. With the tutorials in this hands-on guide, you'll learn how to use the essential R tools you need to know to analyze data, including data types and programming concepts. The second half of Learning R shows you real data analysis in action by covering everything from importing data to publishing your results. Each chapter in the book includes a quiz on what you've learned, and concludes with exercises, most of which involve writing R code. Write a simple R program, and discover what the language can do Use data types such as vectors, arrays, lists, data frames, and strings Execute code conditionally or repeatedly with branches and loops Apply R add-on packages, and package your own work for others Learn how to clean data you import from a variety of sources Understand data through visualization and summary statistics Use statistical models to pass quantitative judgments about data and make predictions Learn what to do when things go wrong while writing data analysis code

Designing Data Visualizations - Noah Iliinsky 2011-09-16

Data visualization is an efficient and effective medium for communicating large amounts of information, but the design process can often seem like an unexplainable creative endeavor. This concise book aims to demystify

Downloaded from hoekstratruck.com on by guest

the design process by showing you how to use a linear decision-making process to encode your information visually. Delve into different kinds of visualization, including infographics and visual art, and explore the influences at work in each one. Then learn how to apply these concepts to your design process. Learn data visualization classifications, including explanatory, exploratory, and hybrid Discover how three fundamental influences—the designer, the reader, and the data—shape what you create Learn how to describe the specific goal of your visualization and identify the supporting data Decide the spatial position of your visual entities with axes Encode the various dimensions of your data with appropriate visual properties, such as shape and color See visualization best practices and suggestions for encoding various specific data types

Downloaded from hoekstratruck.com on by guest