Pig Tutorial Cloudera

Yeah, reviewing a books **Pig Tutorial Cloudera** could ensue your near associates listings. This is just one of the solutions for you to be successful. As understood, finishing does not recommend that you have astonishing points.

Comprehending as capably as pact even more than new will offer each success. adjacent to, the statement as skillfully as acuteness of this Pig Tutorial Cloudera can be taken as well as picked to act.

Cloudera Administration Handbook - Rohit Menon 2014-07-18
An easy-to-follow Apache Hadoop administrator’s guide filled with practical screenshots and explanations for each step and configuration. This book is great for administrators interested in setting up and managing a large Hadoop cluster. If you are an administrator, or want to be an administrator, and you are ready to build and maintain a production-level cluster running CDH5, then this book is for you.
Hadoop Practice Guide - Jisha Mariam Jose 2019-08-19
This book is a complete practical approach for Hadoop lovers. It is mainly aimed at beginners who want to have a hands-on experience with Hadoop and its ecosystem. Its simplicity and step-by-step explanation will help students and other readers in the computer science industry to use this book as a reference manual. The book has been divided into various chapters that cover Hadoop installation, Summary on Hadoop core components, General commands in Hadoop with examples, SQOOP-import & export commands with verification steps, Pig Latin Commands, Analysis using Pig Latin, Pig Script examples, HiveQL Queries and expected outputs and HBase with CRUD operations. In short, this book is a guide for programmers and non-programmers to begin their projects in Hadoop. It is also suitable as a reference manual for students and professionals who are new to the Hadoop Ecosystems.
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
Data-intensive Text Processing with MapReduce - Jimmy Lin 2010
Our world is being revolutionized by data-driven methods: access to large amounts of data has generated new insights and opened exciting new opportunities in commerce, science, and computing applications. Processing the enormous quantities of data necessary for these advances requires large clusters, making distributed computing paradigms more crucial than ever. MapReduce is a programming model for expressing distributed computations on massive datasets and an execution framework for large-scale data processing on clusters of commodity servers. The programming model provides an easy-to-understand abstraction for designing scalable algorithms, while the execution framework transparently handles many system-level details, ranging from scheduling to synchronization to fault tolerance. This book focuses on MapReduce algorithm design, with an emphasis on text processing algorithms common in natural language processing, information retrieval, and machine learning. We introduce the notion of MapReduce design patterns, which represent general reusable solutions to commonly occurring problems across a variety of problem domains. This book not only intends to help the reader "think in MapReduce", but also discusses limitations of the programming model as well. This volume is a printed version of a work that appears in the Synthesis Digital Library of Engineering and Computer Science. Synthesis Lectures provide concise, original presentations of important research and development topics, published quickly, in digital and print formats. For more information visit www.morganclaypool.com
Big Data and Hadoop - VK Jain 2017-01-01
This book introduces you to the Big Data processing techniques addressing but not limited to various BI (business intelligence) requirements, such as reporting, batch analytics, online analytical processing (OLAP), data mining and Warehousing, and predictive analytics. The book has been written on IBMs Platform of Hadoop framework. IBM Infosphere BigInsight has the highest amount of tutorial matter available free of cost on Internet which makes it easy to acquire proficiency in this technique. This therefore becomes highly vunerable coaching materials in easy to learn steps. The book optimally provides the courseware as per MCA and M. Tech Level Syllabi of most of the Universities. All components of big Data Platform like Jaql, Hive Pig, Sqoop, Flume , Hadoop Streaming, Oozie: HBase, HDFS, FlumeNG, Whirr, Cloudera, Fuse , Zookeeper and Mahout: Machine learning for Hadoop has been discussed in sufficient Detail with hands on Exercises on each.
SQL Server 2017 Integration Services Cookbook - Christian Cote 2017-06-30
Harness the power of SQL Server 2017 Integration Services to build your data integration solutions with ease About This Book Acquaint yourself with all the newly introduced features in SQL Server 2017 Integration Services Program and extend your packages to enhance their functionality This detailed, step-by-step guide covers everything you need to develop efficient data integration and data transformation solutions for your organization Who This Book Is For This book is ideal for software engineers, DW/ETL architects, and ETL developers who need to create a new, or enhance an existing, ETL implementation with SQL Server 2017 Integration Services. This book would also be good for individuals who develop ETL solutions that use SSIS and are keen to learn the new features and capabilities in SSIS 2017. What You Will Learn Understand the key components of an ETL solution using SQL Server 2016-2017 Integration Services Design the architecture of a modern ETL solution Have a good knowledge of the new capabilities and features added to Integration Services Implement ETL solutions using Integration Services for both on-premises and Azure data Improve the performance and scalability of an ETL solution Enhance the ETL solution using a custom framework Be able to work on the ETL solution with many other developers and have common design paradigms or techniques Effectively use scripting to solve complex data issues In Detail SQL Server Integration Services is a tool that facilitates data extraction, consolidation, and loading options (ETL), SQL Server coding enhancements, data warehousing, and customizations. With the help of the recipes in this book, you'll gain complete hands-on experience of SSIS 2017 as well as the 2016 new features, design and development improvements including SCD, Tuning, and Customizations. At the start, you'll learn to install and set up SSIS as well other SQL Server resources to make optimal use of this Business Intelligence tools. We'll begin by taking you through the new features in SSIS 2016/2017 and implementing the necessary features to get a modern scalable ETL solution that fits the modern data warehouse. Through the course of chapters, you will learn how to design and build SSIS data warehouses packages using SQL Server Data Tools.

Additionally, you'll learn to develop SSIS packages designed to maintain a data warehouse using the Data Flow and other control flow tasks. You'll also be demonstrated many recipes on cleansing data and how to get the end result after applying different transformations. Some real-world scenarios that you might face are also covered and how to handle various issues that you might face when designing your packages. At the end of this book, you'll get to know all the key concepts to perform data integration and transformation. You'll have explored on-premises Big Data integration processes to create a classic data warehouse, and will know how to extend the toolbox with custom tasks and transforms. Style and approach This cookbook follows a problem-solution approach and tackles all kinds of data integration scenarios by using the capabilities of SQL Server 2016 Integration Services. This book is well supplemented with screenshots, tips, and tricks. Each recipe focuses on a particular task and is written in a very easy-to-follow manner.

Hadoop: The Definitive Guide - Tom White 2015-03-25

Get ready to unlock the power of your data. With the fourth edition of this comprehensive guide, youâ??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. Using Hadoop 2 exclusively, author Tom White presents new chapters on YARN and several Hadoop-related projects such as Parquet, Flume, Crunch, and Spark. Youâ??ll learn about recent changes to Hadoop, and explore new case studies on Hadoopâ??s role in healthcare systems and genomics data processing. Learn fundamental components such as MapReduce, HDFS, and YARN Explore MapReduce in depth, including steps for developing applications with it Set up and maintain a Hadoop cluster running HDFS and MapReduce on YARN Learn two data formats: Avro for data serialization and Parquet for nested data Use data ingestion tools such as Flume (for streaming data) and Sqoop (for bulk data transfer) Understand how high-level data processing tools like Pig, Hive, Crunch, and Spark work with Hadoop Learn the HBase distributed database and the ZooKeeper distributed configuration service

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.

Hadoop Operations - Eric Sammer 2012-09-26

If you've been asked to maintain large and complex Hadoop clusters, this book is a must. Demand for operations-specific material has skyrocketed now that Hadoop is becoming the de facto standard for truly large-scale data processing in the data center. Eric Sammer, Principal Solution Architect at Cloudera, shows you the particulars of running Hadoop in production, from planning, installing, and configuring the system to providing ongoing maintenance. Rather than run through all possible scenarios, this pragmatic operations guide calls out what works, as demonstrated in critical deployments. Get a high-level overview of HDFS and MapReduce: why they exist and how they work Plan a Hadoop deployment, from hardware and OS selection to network requirements Learn setup and configuration details with a list of critical properties Manage resources by sharing a cluster across multiple groups Get a runbook of the most common cluster maintenance tasks Monitor Hadoop clusters—and learn troubleshooting with the help of real-world war stories Use basic tools and techniques to handle backup and catastrophic failure

HADOOP - TOM. WHITE 2015

Big Data Made Easy - Michael Frampton 2014-12-31

Many corporations are finding that the size of their data sets are outgrowing the capability of their systems to store and process them. The data is becoming too big to manage and use with traditional tools. The solution: implementing a big data system. As Big Data Made Easy: A Working Guide to the Complete Hadoop Toolset shows, Apache Hadoop offers a scalable, fault-tolerant system for storing and processing data in parallel. It has a very rich toolset that allows for storage (Hadoop), configuration (YARN and ZooKeeper), collection (Nutch and Solr), processing (Storm, Pig, and Map Reduce), scheduling (Oozie), moving (Sqoop and Avro), monitoring (Chukwa, Ambari, and Hue), testing (Big Top), and analysis (Hive). The problem is that the Internet offers IT pros wading into big data many versions of the truth and some outright falsehoods born of ignorance. What is needed is a book just like this one: a wide-ranging but easily understood set of instructions to explain where to get Hadoop tools, what they can do, how to install them, how to configure them, how to integrate them, and how to use them successfully. And you need an expert who has worked in this area for a decade—someone just like author and big data expert Mike Frampton. Big Data Made Easy approaches the problem of managing massive data sets from a systems perspective, and it explains the roles for each project (like architect and tester, for example) and shows how the Hadoop toolset can be used at each system stage. It explains, in an easily understood manner and through numerous examples, how to use each tool. The book also explains the sliding scale of tools available depending upon data size and when and how to use them. Big Data Made Easy shows developers and architects, as well as testers and project managers, how to: Store big data Configure big data Process big data Schedule processes Move data among SQL and NoSQL systems Monitor data Perform big data analytics Report on big data processes and projects Test big data systems Big Data Made Easy also explains the best part, which is that this toolset is free. Anyone can download it and—with the help of this book—start to use it within a day. With the skills this book will teach you under your belt, you will add value to your company or client immediately, not to mention your career.

Cloud Analytics with Google Cloud Platform - Sanket Thodge 2018-04-10

Combine the power of analytics and cloud computing for faster and efficient insights Key Features Master the concept of analytics on the cloud: and how organizations are using it Learn the design considerations and while applying a cloud analytics solution Design an end-to-end analytics pipeline on the cloud Book Description With the ongoing data explosion, more and more organizations all over the world are slowly migrating their infrastructure to the cloud. These cloud platforms also provide their distinct analytics services to help you get faster insights from your data. This book will give you an introduction to the concept of analytics on the cloud, and the different cloud services popularly used for processing and analyzing data. If you're planning to adopt the cloud analytics model for your business, this book will help you understand the design and business considerations to be kept in mind, and choose the best tools and alternatives for analytics, based on your requirements. The chapters in this book will take you through the 70+ services available in Google Cloud Platform and their implementation for practical purposes. From ingestion to processing your data, this book contains best practices on building an end-to-end analytics pipeline on the cloud by leveraging popular concepts such as machine learning and deep learning. By the end of this book, you will have a better understanding of cloud analytics as a concept as well as a practical know-how of its implementation What you will learn Explore the basics of cloud analytics and the major cloud solutions Learn how organizations are using cloud analytics to improve the ROI Explore the design considerations while adopting cloud services Work with the ingestion and storage tools of GCP such as Cloud Pub/Sub Process your data with tools such as Cloud Dataproc, BigQuery, etc Over 70 GCP tools to build an analytics engine for cloud analytics Implement machine learning and other AI techniques on GCP Who this book is for This book is targeted at CIOs, CTOs, and even analytics professionals looking for various alternatives to implement their analytics pipeline on the cloud. Data professionals looking to get started with cloud-based analytics will also find this book useful. Some basic exposure to cloud platforms such as GCP will be helpful, but not mandatory.

Field Guide to Hadoop - Kevin Sitto 2015-03-02

Annotation IT Managers, developers, data analysts, system architects, and similar technical workers are now encountering the largest and most disruptive change in their profession since the ascendancy of the relational database in early 1980s. You hear that NoSQL and Big Data Analytics are about to replace the systems and skills you now own and possess, but there's often no easy way to make that transition.To exacerbate the issue, the transition may not be gradual, but forced on you by a new project in your enterprisenamely, Hadoopthat will immediately require new ways of thinking, new tools, and new techniques.This book helps you understand the components of the Hadoop ecosystem and how they relate to each other. You'll discover how to get started on that project in an efficient manner that lays out the possibilities. The authors suggest a path and resources that will guide you on their journey from the status quo to the Brave New World you face.

Designing Data-Intensive Applications - Martin Kleppmann 2017-03-16

Data is at the center of many challenges in system design today. Difficult issues need to be figured out, such as scalability, consistency, reliability, efficiency, and maintainability. In addition, we have an overwhelming variety of tools, including relational databases, NoSQL datastores, stream or batch processors, and message brokers. What are the right choices for your application? How do you make sense of all these buzzwords? In this

practical and comprehensive guide, author Martin Kleppmann helps you navigate this diverse landscape by examining the pros and cons of various technologies for processing and storing data. Software keeps changing, but the fundamental principles remain the same. With this book, software engineers and architects will learn how to apply those ideas in practice, and how to make full use of data in modern applications. Peer under the hood of the systems you already use, and learn how to use and operate them more effectively Make informed decisions by identifying the strengths and weaknesses of different tools Navigate the trade-offs around consistency, scalability, fault tolerance, and complexity Understand the distributed systems research upon which modern databases are built Peek behind the scenes of major online services, and learn from their architectures

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.

Impala in Action - Ricky Saltzer 2015-04-07

Hadoop queries in Pig or Hive can be too slow for real-time data analysis. Impala, an ultra-speedy query engine from Cloudera, supercharges Hadoop by avoiding the typical Map-Reduce overhead and parallelizing queries so that they can run on multiple nodes. This is a big deal for big data, because with Impala, querying Hadoop takes seconds rather than minutes. Impala's dialect is close to standard SQL, and Impala seamlessly accesses HBase and HDFS (Hadoop Distributed File System), allowing considerable freedom in choice of data formats. Impala in Action is a hands-on guide to querying Hadoop using Impala. It starts by comparing Impala to traditional databases and database services on Hadoop. Then it explains Impala's SQL dialect and the basics of data access. Next, it tackles data visualization tasks and provides techniques for securing Impala with Apache Sentry. The book also shows how to embed Impala queries in a Java client and how to connect to JDBC and ODBC clients. Advanced readers will appreciate the deep dive into Impala's architecture and the practical insights into the issues complicated configurations and complex queries can cause. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications.

Hadoop For Dummies - Dirk deRoos 2014-03-21

Let Hadoop For Dummies help harness the power of yourdata and rein in the information overload Big data has become big business, and companies and organizationsof all sizes are struggling to find ways to retrieve valuableinformation from their massive data sets with becoming overwhelmed.Enter Hadoop and this easy-to-understand For Dummiesguide. Hadoop For Dummies helps readers understand thevalue of big data, make a business case for using Hadoop, navigatethe Hadoop ecosystem, and build and manage Hadoop applications andclusters. Explains the origins of Hadoop, its economic benefits, and itsfunctionality and practical applications Helps you find your way around the Hadoop ecosystem, programMapReduce, utilize design patterns, and get your Hadoop cluster upand running quickly and easily Details how to use Hadoop applications for data mining, webanalytics and personalization, large-scale text processing, datascience, and problem-solving Shows you how to improve the value of your Hadoop cluster,maximize your investment in Hadoop, and avoid common pitfalls whenbuilding your Hadoop cluster From programmers challenged with building and maintainingaffordable, scaleable data systems to administrators who must dealwith huge volumes of information effectively and efficiently, thishow-to has something to help you with Hadoop.

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.

Practical Hive - Scott Shaw 2016-08-27

Dive into the world of SQL on Hadoop and get the most out of your Hive data warehouses. This book is your go-to resource for using Hive: authors Scott Shaw, Ankur Gupta, David Kjerrumgaard, and Andreas Francois Vermeulen take you through learning HiveQL, the SQL-like language specific to Hive, to analyze, export, and massage the data stored across your Hadoop environment. From deploying Hive on your hardware or virtual machine and setting up its initial configuration to learning how Hive interacts with Hadoop, MapReduce, Tez and other big data technologies, Practical Hive gives you a detailed treatment of the software. In addition, this book discusses the value of open source software, Hive performance tuning, and how to leverage semi-structured and unstructured data. What You Will Learn Install and configure Hive for new and existing datasets Perform DDL operations Execute efficient DML operations Use tables, partitions, buckets, and user-defined functions Discover performance tuning tips and Hive best practices Who This Book Is For Developers, companies, and professionals who deal with large amounts of data and could use software that can efficiently manage large volumes of input. It is assumed that readers have the ability to work with SQL.

Professional Hadoop Solutions - Boris Lublinsky 2013-09-12

The go-to guidebook for deploying Big Data solutions withHadoop Today's enterprise architects need to understand how the Hadoopframeworks and APIs fit together, and how they can be integrated todeliver real-world solutions. This book is a practical, detailedguide to building and implementing those solutions, with code-levelinstruction in the popular Wrox tradition. It covers storing datawith HDFS and Hbase, processing data with MapReduce, and automatingdata processing with Oozie. Hadoop security, running Hadoop withAmazon Web Services, best practices, and automating Hadoopprocesses in real time are also covered in depth. With in-depth code examples in Java and XML and the latest onrecent additions to the Hadoop ecosystem, this complete resourcealso covers the use of APIs, exposing their inner workings andallowing architects and developers to better leverage and customizethem. The ultimate guide for developers, designers, and architectswho need to build and deploy Hadoop applications Covers storing and processing data with various technologies,automating data processing, Hadoop security, and deliveringreal-time solutions Includes detailed, real-world examples and code-levelguidelines Explains when, why, and how to use these tools effectively Written by a team of Hadoop experts in theprogrammer-to-programmer Wrox style Professional Hadoop Solutions is the reference enterprisearchitects and developers need to maximize the power of Hadoop.

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

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!

Accumulo - Aaron Cordova 2015-07

Get up to speed on Apache Accumulo, the flexible, high-performance key/value store created by the National Security Agency (NSA) and based on Google's BigTable data storage system. Written by former NSA team members, this comprehensive tutorial and reference covers Accumulo architecture, application development, table design, and cell-level security. With clear information on system administration, performance tuning, and best practices, this book is ideal for developers seeking to write Accumulo applications, administrators charged with installing and maintaining Accumulo, and other professionals interested in what Accumulo has to offer. You will find everything you need to use this system fully. Get a high-level introduction to Accumulo's architecture and data model Take a rapid tour through single- and multiple-node installations, data ingest, and query Learn how to write Accumulo applications for several use cases, based on examples Dive into Accumulo internals, including information not available in the documentation Get detailed information for installing, administering, tuning, and measuring performance Learn best practices based on successful implementations in the field Find answers to common questions that every new Accumulo user asks

Hadoop Application Architectures - Mark Grover 2015-06-30

Get expert guidance on architecting end-to-end data management solutions with Apache Hadoop. While many sources explain how to use various components in the Hadoop ecosystem, this practical book takes you through architectural considerations necessary to tie those components together into a complete tailored application, based on your particular use case. To reinforce those lessons, the book's second section provides detailed examples of architectures used in some of the most commonly found Hadoop applications. Whether you're designing a new Hadoop application, or planning to integrate Hadoop into your existing data infrastructure, Hadoop Application Architectures will skillfully guide you through the process. This book covers: Factors to consider when using Hadoop to store and model data Best practices for moving data in and out of the system Data processing frameworks, including MapReduce, Spark, and Hive Common Hadoop processing patterns, such as removing duplicate records and using windowing analytics Giraph, GraphX, and other tools for large graph processing on Hadoop Using workflow orchestration and scheduling tools such as Apache Oozie Near-real-time stream processing with Apache Storm, Apache Spark Streaming, and Apache Flume Architecture examples for clickstream analysis, fraud detection, and data warehousing

Essentials of Business Analytics - Bhimasankaram Pochiraju 2019-07-10

This comprehensive edited volume is the first of its kind, designed to serve as a textbook for long-duration business analytics programs. It can also be used as a guide to the field by practitioners. The book has contributions from experts in top universities and industry. The editors have taken extreme care to ensure continuity across the chapters. The material is organized into three parts: A) Tools, B) Models and C) Applications. In Part A, the tools used by business analysts are described in detail. In Part B, these tools are applied to construct models used to solve business problems. Part C contains detailed applications in various functional areas of business and several case studies. Supporting material can be found in the appendices that develop the pre-requisites for the main text. Every chapter has a business orientation. Typically, each chapter begins with the description of business problems that are transformed into data questions; and methodology is developed to solve these questions. Data analysis is conducted using widely used software, the output and results are clearly explained at each stage of development. These are finally transformed into a business solution. The companion website provides examples, data sets and sample code for each chapter.

MapReduce Design Patterns - Donald Miner 2012-11-21

Until now, design patterns for the MapReduce framework have been scattered among various research papers, blogs, and books. This handy guide brings together a unique collection of valuable MapReduce patterns that will save you time and effort regardless of the domain, language, or development framework you're using. Each pattern is explained in context, with pitfalls and caveats clearly identified to help you avoid common design mistakes when modeling your big data architecture. This book also provides a complete overview of MapReduce that explains its origins and implementations, and why design patterns are so important. All code examples are written for Hadoop. Summarization patterns: get a top-level view by summarizing and grouping data Filtering patterns: view data subsets such as records generated from one user Data organization patterns: reorganize data to work with other systems, or to make MapReduce analysis easier Join patterns: analyze different datasets together to discover interesting relationships Metapatterns: piece together several patterns to solve multi-stage problems, or to perform several analytics in the same job Input and output patterns: customize the way you use Hadoop to load or store data "A clear exposition of MapReduce programs for common data processing patterns—this book is indespensible for anyone using Hadoop." --Tom White, author of Hadoop: The Definitive Guide

The Data Science Framework - Juan J. Cuadrado-Gallego 2020-10-01

This edited book first consolidates the results of the EU-funded EDISON project (Education for Data Intensive Science to Open New science frontiers), which developed training material and information to assist educators, trainers, employers, and research infrastructure managers in identifying, recruiting and inspiring the data science professionals of the future. It then deepens the presentation of the information and knowledge gained to allow for easier assimilation by the reader. The contributed chapters are presented in sequence, each chapter picking up from the end point of the previous one. After the initial book and project overview, the chapters present the relevant data science competencies and body of knowledge, the model curriculum required to teach the required foundations, profiles of professionals in this domain, and use cases and applications. The text is supported with appendices on related process models. The book can be used to develop new courses in data science, evaluate existing modules and courses, draft job descriptions, and plan and design efficient data-intensive research teams across scientific disciplines.

Programming Hive - Edward Capriolo 2012-09-26

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

Professional NoSQL - Shashank Tiwari 2011-08-31

A hands-on guide to leveraging NoSQL databases NoSQL databases are an efficient and powerful tool for storing and manipulating vast quantities of data. Most NoSQL databases scale well as data grows. In addition, they are often malleable and flexible enough to accommodate semi-structured and sparse data sets. This comprehensive hands-on guide presents fundamental concepts and practical solutions for getting you ready to use NoSQL databases. Expert author Shashank Tiwari begins with a helpful introduction on the subject of NoSQL, explains its characteristics and typical uses, and looks at where it fits in the application stack. Unique insights help you choose which NoSQL solutions are best for solving your specific data storage needs. Professional NoSQL: Demystifies the concepts

that relate to NoSQL databases, including column-family oriented stores, key/value databases, and document databases. Delves into installing and configuring a number of NoSQL products and the Hadoop family of products. Explains ways of storing, accessing, and querying data in NoSQL databases through examples that use MongoDB, HBase, Cassandra, Redis, CouchDB, Google App Engine Datastore and more. Looks at architecture and internals. Provides guidelines for optimal usage, performance tuning, and scalable configurations. Presents a number of tools and utilities relating to NoSQL, distributed platforms, and scalable processing, including Hive, Pig, RRDtool, Nagios, and more.

Apache Hadoop 3 Quick Start Guide - Hrishikesh Vijay Karambelkar 2018-10-31

A fast paced guide that will help you learn about Apache Hadoop 3 and its ecosystem Key Features Set up, configure and get started with Hadoop to get useful insights from large data sets Work with the different components of Hadoop such as MapReduce, HDFS and YARN Learn about the new features introduced in Hadoop 3 Book Description Apache Hadoop is a widely used distributed data platform. It enables large datasets to be efficiently processed instead of using one large computer to store and process the data. This book will get you started with the Hadoop ecosystem, and introduce you to the main technical topics, including MapReduce, YARN, and HDFS. The book begins with an overview of big data and Apache Hadoop. Then, you will set up a pseudo Hadoop development environment and a multi-node enterprise Hadoop cluster. You will see how the parallel programming paradigm, such as MapReduce, can solve many complex data processing problems. The book also covers the important aspects of the big data software development lifecycle, including quality assurance and control, performance, administration, and monitoring. You will then learn about the Hadoop ecosystem, and tools such as Kafka, Sqoop, Flume, Pig, Hive, and HBase. Finally, you will look at advanced topics, including real time streaming using Apache Storm, and data analytics using Apache Spark. By the end of the book, you will be well versed with different configurations of the Hadoop 3 cluster. What you will learn Store and analyze data at scale using HDFS, MapReduce and YARN Install and configure Hadoop 3 in different modes Use Yarn effectively to run different applications on Hadoop based platform Understand and monitor how Hadoop cluster is managed Consume streaming data using Storm, and then analyze it using Spark Explore Apache Hadoop ecosystem components, such as Flume, Sqoop, HBase, Hive, and Kafka Who this book is for Aspiring Big Data professionals who want to learn the essentials of Hadoop 3 will find this book to be useful. Existing Hadoop users who want to get up to speed with the new features introduced in Hadoop 3 will also benefit from this book. Having knowledge of Java programming will be an added advantage.

Getting Started with Impala - John Russell 2014-09-25

Learn how to write, tune, and port SQL queries and other statements for a Big Data environment, using Impala—the massively parallel processing SQL query engine for Apache Hadoop. The best practices in this practical guide help you design database schemas that not only interoperate with other Hadoop components, and are convenient for administers to manage and monitor, but also accommodate future expansion in data size and evolution of software capabilities. Written by John Russell, documentation lead for the Cloudera Impala project, this book gets you working with the most recent Impala releases quickly. Ideal for database developers and business analysts, the latest revision covers analytics functions, complex types, incremental statistics, subqueries, and submission to the Apache incubator. Getting Started with Impala includes advice from Cloudera's development team, as well as insights from its consulting engagements with customers. Learn how Impala integrates with a wide range of Hadoop components Attain high performance and scalability for huge data sets on production clusters Explore common developer tasks, such as porting code to Impala and optimizing performance Use tutorials for working with billion-row tables, date- and time-based values, and other techniques Learn how to transition from rigid schemas to a flexible model that evolves as needs change Take a deep dive into joins and the roles of statistics

From Visual Surveillance to Internet of Things - Lavanya Sharma 2019-10-16

From Visual Surveillance to Internet of Things: Technology and Applications is an invaluable resource for students, academicians and researchers to explore the utilization of Internet of Things with visual surveillance and its underlying technologies in different application areas. Using a series of present and future applications – business insights, indoor-outdoor securities, smart grids, human detection and tracking, intelligent traffic monitoring, e-health department and many more – this book will support readers to obtain a deeper knowledge in implementing IoT with visual surveillance. The book offers comprehensive coverage of the most essential topics, including: The rise of machines and communications to IoT (3G, 5G) Tools and technologies of IoT with visual surveillance IoT with visual surveillance for real-time applications IoT architectures Challenging issues and novel solutions for realistic applications Mining and tracking of motion-based object data Image processing and analysis into the unified framework to understand both IOT and computer vision applications This book will be an ideal resource for IT professionals, researchers, under- or post-graduate students, practitioners, and technology developers who are interested in gaining a deeper knowledge in implementing IoT with visual surveillance, critical applications domains, technologies, and solutions to handle relevant challenges. Dr. Lavanya Sharma is an Assistant Professor in the Amity Institute of Information Technology at Amity University UP, Noida, India. She is a recipient of several prestigious awards during her academic career. She is an active nationally-recognized researcher who has published numerous papers in her field. She has contributed as an Organizing Committee member and session chair at Springer and IEEE conferences. Prof. Pradeep K. Garg worked as a Vice Chancellor, Uttarakhand Technical University, Dehradun. Presently he is working in the department of Civil Engineering, IIT Roorkee as a professor. Prof. Garg has published more than 300 technical papers in national and international conferences and journals. He has completed 26 research projects funded by various government agencies, guided 27 PhD candidates, and provided technical services to 84 consultancy projects on various aspects of Civil Engineering.

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

The data lake is a daring new approach for harnessing the power of big data technology and providing convenient self-service capabilities. But is it right for your company? This book is based on discussions with practitioners and executives from more than a hundred organizations, ranging from data-driven companies such as Google, LinkedIn, and Facebook, to governments and traditional corporate enterprises. You'll learn what a data lake is, why enterprises need one, and how to build one successfully with the best practices in this book. Alex Gorelik, CTO and founder of Waterline Data, explains why old systems and processes can no longer support data needs in the enterprise. Then, in a collection of essays about data lake implementation, you'll examine data lake initiatives, analytic projects, experiences, and best practices from data experts working in various industries. Get a succinct introduction to data warehousing, big data, and data science Learn various paths enterprises take to build a data lake Explore how to build a self-service model and best practices for providing analysts access to the data Use different methods for architecting your data lake Discover ways to implement a data lake from experts in different industries

Expert Hadoop 2 Administration - Sam R. Alapati 2016-11-29

This is the eBook of the printed book and may not include any media, website access codes, or print supplements that may come packaged with the bound book. The Comprehensive, Up-to-Date Apache Hadoop Administration Handbook and Reference "Sam Alapati has worked with production Hadoop clusters for six years. His unique depth of experience has enabled him to write the go-to resource for all administrators looking to spec, size, expand, and secure production Hadoop clusters of any size." —Paul Dix, Series Editor In Expert Hadoop® Administration, leading Hadoop administrator Sam R. Alapati brings together authoritative knowledge for creating, configuring, securing, managing, and optimizing production Hadoop clusters in any environment. Drawing on his experience with large-scale Hadoop administration, Alapati integrates action-oriented advice with carefully researched explanations of both problems and solutions. He covers an unmatched range of topics and offers an unparalleled collection of realistic examples. Alapati demystifies complex Hadoop environments, helping you understand exactly what happens behind the scenes when you administer your cluster. You'll gain unprecedented insight as you walk through building clusters from scratch and configuring high availability, performance, security, encryption, and other key attributes. The high-value administration skills you learn here will be indispensable no matter what Hadoop distribution you use or what Hadoop applications you run. Understand Hadoop's architecture from an administrator's standpoint Create simple and fully distributed clusters Run MapReduce and Spark applications in a Hadoop cluster Manage and protect Hadoop data and high availability Work with HDFS commands, file permissions, and storage management Move data, and use YARN to allocate resources and schedule jobs Manage job workflows with Oozie and Hue Secure, monitor, log, and optimize Hadoop Benchmark and

troubleshoot Hadoop

Hadoop in Practice - Alex Holmes 2014-10-12

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

Enabling the New Era of Cloud Computing: Data Security, Transfer, and Management - Shen, Yushi 2013-11-30

Cloud computing is becoming the next revolution in the IT industry; providing central storage for internet data and services that have the potential to bring data transmission performance, security and privacy, data deluge, and inefficient architecture to the next level. Enabling the New Era of Cloud Computing: Data Security, Transfer, and Management discusses cloud computing as an emerging technology and its critical role in the IT industry upgrade and economic development in the future. This book is an essential resource for business decision makers, technology investors, architects and engineers, and cloud consumers interested in the cloud computing future.

Hadoop Beginner's Guide - Garry Turkington 2013-02-22

Data is arriving faster than you can process it and the overall volumes keep growing at a rate that keeps you awake at night. Hadoop can help you tame the data beast. Effective use of Hadoop however requires a mixture of programming, design, and system administration skills. "Hadoop Beginner's Guide" removes the mystery from Hadoop, presenting Hadoop and related technologies with a focus on building working systems and getting the job done, using cloud services to do so when it makes sense. From basic concepts and initial setup through developing applications and keeping the system running as the data grows, the book gives the understanding needed to effectively use Hadoop to solve real world problems. Starting with the basics of installing and configuring Hadoop, the book explains how to develop applications, maintain the system, and how to use additional products to integrate with other systems. While learning different ways to develop applications to run on Hadoop the book also covers tools such as Hive, Sqoop, and Flume that show how Hadoop can be integrated with relational databases and log collection. In addition to examples on Hadoop clusters on Ubuntu uses of cloud services such as Amazon, EC2 and Elastic MapReduce are covered.

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.

Apache Hadoop YARN - Arun C. Murthy 2014

"Apache Hadoop is helping drive the Big Data revolution. Now, its data processing has been completely overhauled: Apache Hadoop YARN provides resource management at data center scale and easier ways to create distributed applications that process petabytes of data. And now in Apache HadoopTM YARN, two Hadoop technical leaders show you how to develop new applications and adapt existing code to fully leverage these revolutionary advances." -- From the Amazon

Big Data Analytics and Computing for Digital Forensic Investigations - Suneeta Satpathy 2020-03-17

Digital forensics has recently gained a notable development and become the most demanding area in today’s information security requirement. This book investigates the areas of digital forensics, digital investigation and data analysis procedures as they apply to computer fraud and cybercrime, with the main objective of describing a variety of digital crimes and retrieving potential digital evidence. Big Data Analytics and Computing for Digital Forensic Investigations gives a contemporary view on the problems of information security. It presents the idea that protective mechanisms and software must be integrated along with forensic capabilities into existing forensic software using big data computing tools and techniques. Features Describes trends of digital forensics served for big data and the challenges of evidence acquisition Enables digital forensic investigators and law enforcement agencies to enhance their digital investigation capabilities with the application of data science analytics, algorithms and fusion technique This book is focused on helping professionals as well as researchers to get ready with next-generation security systems to mount the rising challenges of computer fraud and cybercrimes as well as with digital forensic investigations. Dr Suneeta Satpathy has more than ten years of teaching experience in different subjects of the Computer Science and Engineering discipline. She is currently working as an associate professor in the Department of Computer Science and Engineering, College of Bhubaneswar, affiliated with Biju Patnaik University and Technology, Odisha. Her research interests include computer forensics, cybersecurity, data fusion, data mining, big data analysis and decision mining. Dr Sachi Nandan Mohanty is an associate professor in the Department of Computer Science and Engineering at ICFAI Tech, ICFAI Foundation for Higher Education, Hyderabad, India. His research interests include data mining, big data analysis, cognitive science, fuzzy decision-making, brain–computer interface, cognition and computational intelligence.