Hadoop The Definitive Guide 3rd Third Edition By White Tom Published By Oreilly Media 2012 Paperback

Cassandra: The Definitive GuideTokyo VernacularHadoop OperationsThe Definitive Guide to MongoDBJavaScript: The Definitive GuideUsing FlumeLearning SparkUsing FlumeHadoop For DummiesHadoop: The Definitive GuideHADOOP IN ACTIONHadoop: The Definitive GuideLearning SQLSpark: The Definitive GuideData Science from ScratchHadoop Application ArchitecturesBig Data with Hadoop MapReduceData AlgorithmsUsing OpenMPAdvances in Computing Systems and ApplicationsSoftware Engineering at GoogleBig Data Analytics with Hadoop 3Introducing Microsoft Azure HDInsightPattern and Data Analysis in Healthcare SettingsCascading Style SheetsKafka: The Definitive GuideApacheThe Data Warehouse ToolkitHBase: The Definitive GuideBig Data Management, Technologies, and ApplicationsPresto: The Definitive GuideHadoop Real-World Solutions CookbookHadoop in PracticeThe Data Warehouse ToolkitProgramming HiveAdvanced Analytics with SparkCassandra: The Definitive GuideExpert Hadoop 2 AdministrationHigh Performance SparkMastering Apache Cassandra

Cassandra: The Definitive Guide

How can you get your data from frontend servers to Hadoop in near real time? With this complete reference guide, you'll learn Flume's rich set of features for collecting, aggregating, and writing large amounts of streaming data to the Hadoop Distributed File System (HDFS), Apache HBase, SolrCloud, Elastic Search, and other systems. Using Flume shows operations engineers how to configure, deploy, and monitor a Flume cluster, and teaches developers how to write Flume plugins and custom components for their specific use-cases. You'll learn about Flume's design and implementation, as well as various features that make it highly scalable, flexible, and reliable. Code examples and exercises are available on GitHub. Learn how Flume provides a steady rate of flow by acting as a buffer between data producers and consumers Dive into key Flume components, including sources that accept data and sinks that write and deliver it Write custom plugins to customize the way Flume receives, modifies, formats, and writes data Explore APIs for sending data to Flume agents from your own applications Plan and deploy Flume in a scalable and flexible way—and monitor your cluster once it's running

Tokyo Vernacular

How can you get your data from frontend servers to Hadoop in near real time? With this complete reference guide, you'll learn Flume's rich set of features for

collecting, aggregating, and writing large amounts of streaming data to the Hadoop Distributed File System (HDFS), Apache HBase, SolrCloud, Elastic Search, and other systems. Using Flume shows operations engineers how to configure, deploy, and monitor a Flume cluster, and teaches developers how to write Flume plugins and custom components for their specific use-cases. You'll learn about Flume's design and implementation, as well as various features that make it highly scalable, flexible, and reliable. Code examples and exercises are available on GitHub. Learn how Flume provides a steady rate of flow by acting as a buffer between data producers and consumers Dive into key Flume components, including sources that accept data and sinks that write and deliver it Write custom plugins to customize the way Flume receives, modifies, formats, and writes data Explore APIs for sending data to Flume agents from your own applications Plan and deploy Flume in a scalable and flexible way—and monitor your cluster once it's running

Hadoop Operations

Explore big data concepts, platforms, analytics, and their applications using the power of Hadoop 3 Key Features Learn Hadoop 3 to build effective big data analytics solutions on-premise and on cloud Integrate Hadoop with other big data tools such as R, Python, Apache Spark, and Apache Flink Exploit big data using Hadoop 3 with real-world examples Book Description Apache Hadoop is the most

popular platform for big data processing, and can be combined with a host of other big data tools to build powerful analytics solutions. Big Data Analytics with Hadoop 3 shows you how to do just that, by providing insights into the software as well as its benefits with the help of practical examples. Once you have taken a tour of Hadoop 3's latest features, you will get an overview of HDFS, MapReduce, and YARN, and how they enable faster, more efficient big data processing. You will then move on to learning how to integrate Hadoop with the open source tools, such as Python and R, to analyze and visualize data and perform statistical computing on big data. As you get acquainted with all this, you will explore how to use Hadoop 3 with Apache Spark and Apache Flink for real-time data analytics and stream processing. In addition to this, you will understand how to use Hadoop to build analytics solutions on the cloud and an end-to-end pipeline to perform big data analysis using practical use cases. By the end of this book, you will be well-versed with the analytical capabilities of the Hadoop ecosystem. You will be able to build powerful solutions to perform big data analytics and get insight effortlessly. What you will learn Explore the new features of Hadoop 3 along with HDFS, YARN, and MapReduce Get well-versed with the analytical capabilities of Hadoop ecosystem using practical examples Integrate Hadoop with R and Python for more efficient big data processing Learn to use Hadoop with Apache Spark and Apache Flink for real-time data analytics Set up a Hadoop cluster on AWS cloud Perform big data analytics on AWS using Elastic Map Reduce Who this book is for Big Data Analytics with Hadoop 3 is for you if you are looking to build high-performance analytics

solutions for your enterprise or business using Hadoop 3’s powerful features, or you’re new to big data analytics. A basic understanding of the Java programming language is required.

The Definitive Guide to MongoDB

This book gathers selected papers presented at the 3rd Conference on Computing Systems and Applications (CSA’2018), held at the Ecole Militaire Polytechnique, Algiers, Algeria on April 24–25, 2018. The CSA’2018 constitutes a leading forum for exchanging, discussing and leveraging modern computer systems technology in such varied fields as: data science, computer networks and security, information systems and software engineering, and computer vision. The contributions presented here will help promote and advance the adoption of computer science technologies in industrial, entertainment, social, and everyday applications. Though primarily intended for students, researchers, engineers and practitioners working in the field, it will also benefit a wider audience interested in the latest developments in the computer sciences.

JavaScript: The Definitive Guide

"This book discusses the exponential growth of information size and the innovative

methods for data capture, storage, sharing, and analysis for big data"--Provided by publisher.

Using Flume

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

Learning Spark

Mastering Apache Cassandra is a practical, hands-on guide with step-by-step instructions. The smooth and easy tutorial approach focuses on showing people how to utilize Cassandra to its full potential.This book is aimed at intermediate Cassandra users. It is best suited for startups where developers have to wear multiple hats: programmer, DevOps, release manager, convincing clients, and handling failures. No prior knowledge of Cassandra is required.

Using Flume

Apache Spark is amazing when everything clicks. But if you haven't seen the performance improvements you expected, or still don't feel confident enough to use Spark in production, this practical book is for you. Authors Holden Karau and Rachel Warren demonstrate performance optimizations to help your Spark queries run faster and handle larger data sizes, while using fewer resources. Ideal for software engineers, data engineers, developers, and system administrators working with large-scale data applications, this book describes techniques that can reduce data infrastructure costs and developer hours. Not only will you gain a more comprehensive understanding of Spark, you'll also learn how to make it sing. With this book, you'll explore: How Spark SQL's new interfaces improve

performance over SQL's RDD data structure The choice between data joins in Core Spark and Spark SQL Techniques for getting the most out of standard RDD transformations How to work around performance issues in Spark's key/value pair paradigm Writing high-performance Spark code without Scala or the JVM How to test for functionality and performance when applying suggested improvements Using Spark MLlib and Spark ML machine learning libraries Spark's Streaming components and external community packages

Hadoop For Dummies

If you are ready to dive into the MapReduce framework for processing large datasets, this practical book takes you step by step through the algorithms and tools you need to build distributed MapReduce applications with Apache Hadoop or Apache Spark. Each chapter provides a recipe for solving a massive computational problem, such as building a recommendation system. You'll learn how to implement the appropriate MapReduce solution with code that you can use in your projects. Dr. Mahmoud Parsian covers basic design patterns, optimization techniques, and data mining and machine learning solutions for problems in bioinformatics, genomics, statistics, and social network analysis. This book also includes an overview of MapReduce, Hadoop, and Spark. Topics include: Market basket analysis for a large set of transactions Data mining algorithms (K-means, KNN, and Naive Bayes) Using huge genomic data to sequence DNA and RNA Naive

Bayes theorem and Markov chains for data and market prediction Recommendation algorithms and pairwise document similarity Linear regression, Cox regression, and Pearson correlation Allelic frequency and mining DNA Social network analysis (recommendation systems, counting triangles, sentiment analysis)

Hadoop: The Definitive Guide

Today, software engineers need to know not only how to program effectively but also how to develop proper engineering practices to make their codebase sustainable and healthy. This book emphasizes this difference between programming and software engineering. How can software engineers manage a living codebase that evolves and responds to changing requirements and demands over the length of its life? Based on their experience at Google, software engineers Titus Winters and Hyrum Wright, along with technical writer Tom Manshreck, present a candid and insightful look at how some of the world's leading practitioners construct and maintain software. This book covers Google's unique engineering culture, processes, and tools and how these aspects contribute to the effectiveness of an engineering organization. You'll explore three fundamental principles that software organizations should keep in mind when designing, architecting, writing, and maintaining code: How time affects the sustainability of software and how to make your code resilient over time How scale affects the

viability of software practices within an engineering organization What trade-offs a typical engineer needs to make when evaluating design and development decisions

HADOOP IN ACTION

For web developers and other programmers interested in using JavaScript, this bestselling book provides the most comprehensive JavaScript material on the market. The seventh edition represents a significant update, with new information for ECMAScript 2020, and new chapters on language-specific features. JavaScript: The Definitive Guide is ideal for experienced programmers who want to learn the programming language of the web, and for current JavaScript programmers who want to master it.

Hadoop: The Definitive Guide

Let Hadoop For Dummies help harness the power of your data and rein in the information overload Big data has become big business, and companies and organizations of all sizes are struggling to find ways to retrieve valuable information from their massive data sets with becoming overwhelmed. Enter Hadoop and this easy-to-understand For Dummies guide. Hadoop For Dummies

helps readers understand the value of big data, make a business case for using Hadoop, navigate the Hadoop ecosystem, and build and manage Hadoop applications and clusters. Explains the origins of Hadoop, its economic benefits, and its functionality and practical applications Helps you find your way around the Hadoop ecosystem, program MapReduce, utilize design patterns, and get your Hadoop cluster up and running quickly and easily Details how to use Hadoop applications for data mining, web analytics and personalization, large-scale text processing, data science, and problem-solving Shows you how to improve the value of your Hadoop cluster, maximize your investment in Hadoop, and avoid common pitfalls when building your Hadoop cluster From programmers challenged with building and maintaining affordable, scaleable data systems to administrators who must deal with huge volumes of information effectively and efficiently, this how-to has something to help you with Hadoop.

Learning SQL

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

Spark: The Definitive Guide

Every enterprise application creates data, whether it’s log messages, metrics, user activity, outgoing messages, or something else. And how to move all of this data becomes nearly as important as the data itself. If you’re an application architect, developer, or production engineer new to Apache Kafka, this practical guide shows you how to use this open source streaming platform to handle real-time data feeds. Engineers from Confluent and LinkedIn who are responsible for developing Kafka explain how to deploy production Kafka clusters, write reliable event-driven microservices, and build scalable stream-processing applications with this platform. Through detailed examples, you’ll learn Kafka’s design principles, reliability guarantees, key APIs, and architecture details, including the replication

protocol, the controller, and the storage layer. Understand publish-subscribe messaging and how it fits in the big data ecosystem. Explore Kafka producers and consumers for writing and reading messages Understand Kafka patterns and use-case requirements to ensure reliable data delivery Get best practices for building data pipelines and applications with Kafka Manage Kafka in production, and learn to perform monitoring, tuning, and maintenance tasks Learn the most critical metrics among Kafka's operational measurements Explore how Kafka's stream delivery capabilities make it a perfect source for stream processing systems

Data Science from Scratch

In this practical book, four Cloudera data scientists present a set of self-contained patterns for performing large-scale data analysis with Spark. The authors bring Spark, statistical methods, and real-world data sets together to teach you how to approach analytics problems by example. You'll start with an introduction to Spark and its ecosystem, and then dive into patterns that apply common techniques—classification, collaborative filtering, and anomaly detection among others—to fields such as genomics, security, and finance. If you have an entry-level understanding of machine learning and statistics, and you program in Java, Python, or Scala, you'll find these patterns useful for working on your own data applications. Patterns include: Recommending music and the Audioscrobbler data set Predicting forest cover with decision trees Anomaly detection in network traffic

with K-means clustering Understanding Wikipedia with Latent Semantic Analysis Analyzing co-occurrence networks with GraphX Geospatial and temporal data analysis on the New York City Taxi Trips data Estimating financial risk through Monte Carlo simulation Analyzing genomics data and the BDG project Analyzing neuroimaging data with PySpark and Thunder

Hadoop Application Architectures

The Definitive Guide to MongoDB, Second Edition, is updated for the latest version and includes all of the latest MongoDB features, including the aggregation framework introduced in version 2.2 and hashed indexes in version 2.4. MongoDB is the most popular of the "Big Data" NoSQL database technologies, and it's still growing. David Hows from 10gen, along with experienced MongoDB authors Peter Membrey and Eelco Plugge, provide their expertise and experience in teaching you everything you need to know to become a MongoDB pro. What you'll learn Set up MongoDB on all major server platforms, including Windows, Linux, OS X, and cloud platforms like Rackspace, Azure, and Amazon EC2 Work with GridFS and the new aggregation framework Work with your data using non-SQL commands Write applications using either PHP or Python Optimize MongoDB Master MongoDB administration, including replication, replication tagging, and tag-aware sharding Who this book is for Database admins and developers who need to get up to speed on MongoDB and its Big Data, NoSQL approach to dealing with data management.

Table of ContentsPart I: MongoDB Basics Ch. 1: Introduction to MongoDB Ch. 2: Installing MongoDB Ch. 3: The Data Model Ch. 4: Working with Data Ch. 5: GridFS Part II: Developing with MongoDB Ch. 6: PHP and MongoDB Ch. 7: Python and MongoDB Ch. 8: Advanced Queries Part III: Advanced MongoDB with Big Data Ch. 9: Database Administration Ch. 10: Optimization Ch. 11: Replication Ch. 12: Sharding

Big Data with Hadoop MapReduce

Special Features: · Introduction to MapReduce· Examples illustrating ideas in practice· Hadoop's Streaming API· Other related tools, like Pig and Hive About The Book: Hadoop in Action introduces the subject and teaches you how to write programs in the MapReduce style. It starts with a few easy examples and then moves quickly to show Hadoop use in more complex data analysis tasks. Included are best practices and design patterns of MapReduce programming.This book requires basic Java skills. Knowing basic statistical concepts can help with the more advanced examples.

Data Algorithms

Data in all domains is getting bigger. How can you work with it efficiently? Recently

updated for Spark 1.3, this book introduces Apache Spark, the open source cluster computing system that makes data analytics fast to write and fast to run. With Spark, you can tackle big datasets quickly through simple APIs in Python, Java, and Scala. This edition includes new information on Spark SQL, Spark Streaming, setup, and Maven coordinates. Written by the developers of Spark, this book will have data scientists and engineers up and running in no time. You'll learn how to express parallel jobs with just a few lines of code, and cover applications from simple batch jobs to stream processing and machine learning. Quickly dive into Spark capabilities such as distributed datasets, in-memory caching, and the interactive shell Leverage Spark's powerful built-in libraries, including Spark SQL, Spark Streaming, and MLlib Use one programming paradigm instead of mixing and matching tools like Hive, Hadoop, Mahout, and Storm Learn how to deploy interactive, batch, and streaming applications Connect to data sources including HDFS, Hive, JSON, and S3 Master advanced topics like data partitioning and shared variables

Using OpenMP

Over 90 hands-on recipes to help you learn and master the intricacies of Apache Hadoop 2.X, YARN, Hive, Pig, Oozie, Flume, Sqoop, Apache Spark, and Mahout About This Book Implement outstanding Machine Learning use cases on your own analytics models and processes. Solutions to common problems when working with

the Hadoop ecosystem. Step-by-step implementation of end-to-end big data use cases. Who This Book Is For Readers who have a basic knowledge of big data systems and want to advance their knowledge with hands-on recipes. What You Will Learn Installing and maintaining Hadoop 2.X cluster and its ecosystem. Write advanced Map Reduce programs and understand design patterns. Advanced Data Analysis using the Hive, Pig, and Map Reduce programs. Import and export data from various sources using Sqoop and Flume. Data storage in various file formats such as Text, Sequential, Parquet, ORC, and RC Files. Machine learning principles with libraries such as Mahout Batch and Stream data processing using Apache Spark In Detail Big data is the current requirement. Most organizations produce huge amount of data every day. With the arrival of Hadoop-like tools, it has become easier for everyone to solve big data problems with great efficiency and at minimal cost. Grasping Machine Learning techniques will help you greatly in building predictive models and using this data to make the right decisions for your organization. Hadoop Real World Solutions Cookbook gives readers insights into learning and mastering big data via recipes. The book not only clarifies most big data tools in the market but also provides best practices for using them. The book provides recipes that are based on the latest versions of Apache Hadoop 2.X, YARN, Hive, Pig, Sqoop, Flume, Apache Spark, Mahout and many more such ecosystem tools. This real-world-solution cookbook is packed with handy recipes you can apply to your own everyday issues. Each chapter provides in-depth recipes that can be referenced easily. This book provides detailed practices on the latest

technologies such as YARN and Apache Spark. Readers will be able to consider themselves as big data experts on completion of this book. This guide is an invaluable tutorial if you are planning to implement a big data warehouse for your business. Style and approach An easy-to-follow guide that walks you through world of big data. Each tool in the Hadoop ecosystem is explained in detail and the recipes are placed in such a manner that readers can implement them sequentially. Plenty of reference links are provided for advanced reading.

Advances in Computing Systems and Applications

Describes the history of the Web server platform and covers downloading and compiling, configuring and running the program on UNIX, writing specialized modules, and establishing security routines.

Software Engineering at Google

Demonstrates the control and flexibility Cascading Style Sheets bring to Web design, covering selectors and structure, units, text manipulation, colors, backgrounds, borders, visual formatting, and positioning.

Big Data Analytics with Hadoop 3

Using OpenMP discusses hardware developments, describes where OpenMP is applicable, and compares OpenMP to other programming interfaces for shared and distributed memory parallel architectures. It introduces the individual features of OpenMP, provides many source code examples that demonstrate the use and functionality of the language constructs, and offers tips on writing an efficient OpenMP program. It describes how to use OpenMP in full-scale applications to achieve high performance on large-scale architectures, discussing several case studies in detail, and offers in-depth troubleshooting advice. It explains how OpenMP is translated into explicitly multithreaded code, providing a valuable behind-the-scenes account of OpenMP program performance. Finally, Using OpenMP considers trends likely to influence OpenMP development, offering a glimpse of the possibilities of a future OpenMP 3.0 from the vantage point of the current OpenMP 2.5.

Introducing Microsoft Azure HDInsight

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

Pattern and Data Analysis in Healthcare Settings

What could you do with data if scalability wasn't a problem? With this hands-on guide, you'll learn how Apache Cassandra handles hundreds of terabytes of data while remaining highly available across multiple data centers -- capabilities that have attracted Facebook, Twitter, and other data-intensive companies. Cassandra: The Definitive Guide provides the technical details and practical examples you need to assess this database management system and put it to work in a production environment. Author Eben Hewitt demonstrates the advantages of Cassandra's nonrelational design, and pays special attention to data modeling. If you're a developer, DBA, application architect, or manager looking to solve a database scaling issue or future-proof your application, this guide shows you how to harness Cassandra's speed and flexibility. Understand the tenets of Cassandra's column-oriented structure Learn how to write, update, and read Cassandra data Discover how to add or remove nodes from the cluster as your application requires Examine a working application that translates from a relational model to Cassandra's data model Use examples for writing clients in Java, Python, and C# Use the JMX interface to monitor a cluster's usage, memory patterns, and more Tune memory settings, data storage, and caching for better performance

Cascading Style Sheets

The authors provide an understanding of big data and MapReduce by clearly presenting the basic terminologies and concepts. They have employed over 100 illustrations and many worked-out examples to convey the concepts and methods used in big data, the inner workings of MapReduce, and single node/multi-node installation on physical/virtual machines. This book covers almost all the necessary information on Hadoop MapReduce for most online certification exams. Upon completing this book, readers will find it easy to understand other big data processing tools such as Spark, Storm, etc. Ultimately, readers will be able to: • understand what big data is and the factors that are involved • understand the inner workings of MapReduce, which is essential for certification exams • learn the features and weaknesses of MapReduce • set up Hadoop clusters with 100s of physical/virtual machines • create a virtual machine in AWS • write MapReduce with Eclipse in a simple way • understand other big data processing tools and their applications

Kafka: The Definitive Guide

Business and medical professionals rely on large data sets to identify trends or other knowledge that can be gleaned from the collection of it. New technologies

concentrate on data's management, but do not facilitate users' extraction of meaningful outcomes. Pattern and Data Analysis in Healthcare Settings investigates the approaches to shift computing from analysis on-demand to knowledge on-demand. By providing innovative tactics to apply data and pattern analysis, these practices are optimized into pragmatic sources of knowledge for healthcare professionals. This publication is an exhaustive source for policy makers, developers, business professionals, healthcare providers, and graduate students concerned with data retrieval and analysis.

Apache

As data floods into your company, you need to put it to work right away--and SQL is the best tool for the job. With the latest edition of this introductory guide, author Alan Beaulieu helps developers get up to speed with SQL fundamentals for writing database applications, performing administrative tasks, and generating reports. You'll find new chapters on SQL and big data, analytic functions, and working with very large databases. Each chapter presents a self-contained lesson on a key SQL concept or technique using numerous illustrations and annotated examples. Exercises let you practice the skills you learn. Knowledge of SQL is a must for interacting with data. With Learning SQL, you'll quickly discover how to put the power and flexibility of this language to work. Move quickly through SQL basics and several advanced features Use SQL data statements to generate, manipulate, and

retrieve data Create database objects, such as tables, indexes, and constraints with SQL schema statements Learn how datasets interact with queries; understand the importance of subqueries Convert and manipulate data with SQL's built-in functions and use conditional logic in data statements

The Data Warehouse Toolkit

Data science libraries, frameworks, modules, and toolkits are great for doing data science, but they’re also a good way to dive into the discipline without actually understanding data science. In this book, you’ll learn how many of the most fundamental data science tools and algorithms work by implementing them from scratch. If you have an aptitude for mathematics and some programming skills, author Joel Grus will help you get comfortable with the math and statistics at the core of data science, and with hacking skills you need to get started as a data scientist. Today’s messy glut of data holds answers to questions no one’s even thought to ask. This book provides you with the know-how to dig those answers out. Get a crash course in Python Learn the basics of linear algebra, statistics, and probability—and understand how and when they're used in data science Collect, explore, clean, munge, and manipulate data Dive into the fundamentals of machine learning Implement models such as k-nearest Neighbors, Naive Bayes, linear and logistic regression, decision trees, neural networks, and clustering Explore recommender systems, natural language processing, network analysis,

MapReduce, and databases

HBase: The Definitive Guide

Updated new edition of Ralph Kimball's groundbreaking book on dimensional modeling for data warehousing and business intelligence! The first edition of Ralph Kimball's The Data Warehouse Toolkit introduced the industry to dimensional modeling, and now his books are considered the most authoritative guides in this space. This new third edition is a complete library of updated dimensional modeling techniques, the most comprehensive collection ever. It covers new and enhanced star schema dimensional modeling patterns, adds two new chapters on ETL techniques, includes new and expanded business matrices for 12 case studies, and more. Authored by Ralph Kimball and Margy Ross, known worldwide as educators, consultants, and influential thought leaders in data warehousing and business intelligence Begins with fundamental design recommendations and progresses through increasingly complex scenarios Presents unique modeling techniques for business applications such as inventory management, procurement, invoicing, accounting, customer relationship management, big data analytics, and more Draws real-world case studies from a variety of industries, including retail sales, financial services, telecommunications, education, health care, insurance, e-commerce, and more Design dimensional databases that are easy to understand and provide fast query response with The Data Warehouse Toolkit: The Definitive

Guide to Dimensional Modeling, 3rd Edition.

Big Data Management, Technologies, and Applications

Imagine what you could do if scalability wasn't a problem. With this hands-on guide, you'll learn how the Cassandra database management system handles hundreds of terabytes of data while remaining highly available across multiple data centers. This expanded second edition—updated for Cassandra 3.0—provides the technical details and practical examples you need to put this database to work in a production environment. Authors Jeff Carpenter and Eben Hewitt demonstrate the advantages of Cassandra's non-relational design, with special attention to data modeling. If you're a developer, DBA, or application architect looking to solve a database scaling issue or future-proof your application, this guide helps you harness Cassandra's speed and flexibility. Understand Cassandra's distributed and decentralized structure Use the Cassandra Query Language (CQL) and cqlsh—the CQL shell Create a working data model and compare it with an equivalent relational model Develop sample applications using client drivers for languages including Java, Python, and Node.js Explore cluster topology and learn how nodes exchange data Maintain a high level of performance in your cluster Deploy Cassandra on site, in the Cloud, or with Docker Integrate Cassandra with Spark, Hadoop, Elasticsearch, Solr, and Lucene

Presto: The Definitive Guide

Microsoft Azure HDInsight is Microsoft's 100 percent compliant distribution of Apache Hadoop on Microsoft Azure. This means that standard Hadoop concepts and technologies apply, so learning the Hadoop stack helps you learn the HDInsight service. At the time of this writing, HDInsight (version 3.0) uses Hadoop version 2.2 and Hortonworks Data Platform 2.0. In Introducing Microsoft Azure HDInsight, we cover what big data really means, how you can use it to your advantage in your company or organization, and one of the services you can use to do that quickly–specifically, Microsoft's HDInsight service. We start with an overview of big data and Hadoop, but we don't emphasize only concepts in this book–we want you to jump in and get your hands dirty working with HDInsight in a practical way. To help you learn and even implement HDInsight right away, we focus on a specific use case that applies to almost any organization and demonstrate a process that you can follow along with. We also help you learn more. In the last chapter, we look ahead at the future of HDInsight and give you recommendations for self-learning so that you can dive deeper into important concepts and round out your education on working with big data.

Hadoop Real-World Solutions Cookbook

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

Hadoop in Practice

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

The Data Warehouse Toolkit

If you're looking for a scalable storage solution to accommodate a virtually endless amount of data, this book shows you how Apache HBase can fulfill your needs. As the open source implementation of Google's BigTable architecture, HBase scales to billions of rows and millions of columns, while ensuring that write and read performance remain constant. Many IT executives are asking pointed questions about HBase. This book provides meaningful answers, whether you're evaluating this non-relational database or planning to put it into practice right away. Discover how tight integration with Hadoop makes scalability with HBase easier Distribute large datasets across an inexpensive cluster of commodity servers Access HBase with native Java clients, or with gateway servers providing REST, Avro, or Thrift APIs Get details on HBase's architecture, including the storage format, write-ahead log, background processes, and more Integrate HBase with Hadoop's MapReduce framework for massively parallelized data processing jobs Learn how to tune clusters, design schemas, copy tables, import bulk data, decommission nodes, and many other tasks

Programming Hive

Advanced Analytics with Spark

Hadoop: The Definitive Guide helps you harness the power of your data. Ideal for processing large datasets, the Apache Hadoop framework is an open source implementation of the MapReduce algorithm on which Google built its empire. This comprehensive resource demonstrates how to use Hadoop to build reliable, scalable, distributed systems: programmers will find details for analyzing large datasets, and administrators will learn how to set up and run Hadoop clusters. Complete with case studies that illustrate how Hadoop solves specific problems, this book helps you: Use the Hadoop Distributed File System (HDFS) for storing large datasets, and run distributed computations over those datasets using MapReduce Become familiar with Hadoop's data and I/O building blocks for compression, data integrity, serialization, 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 Use Pig, a high-level query language for large-scale data processing Take advantage of HBase, Hadoop's database for structured and semi-structured data Learn ZooKeeper, a toolkit of coordination primitives for building distributed systems If

you have lots of data -- whether it's gigabytes or petabytes -- Hadoop is the perfect solution. Hadoop: The Definitive Guide is the most thorough book available on the subject. "Now you have the opportunity to learn about Hadoop from a master-not only of the technology, but also of common sense and plain talk."-- Doug Cutting, Hadoop Founder, Yahoo!

Cassandra: The Definitive Guide

Perform fast interactive analytics against different data sources using the Presto 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 Presto. Initially developed by Facebook, open source Presto is now used by Netflix, Airbnb, LinkedIn, Twitter, Uber, and many other companies. Matt Fuller, Manfred Moser, and Martin Traverso from Starburst show you how a single Presto query can combine data from multiple sources to allow for analytics across your entire organization. Get started: Explore Presto's use cases and learn about tools that will help you connect to Presto and query data Go deeper: Learn Presto's internal workings, including how to connect to and query data sources with support for SQL statements, operators, functions, and more Put Presto in production: Use this query engine for security and monitoring and with other applications; learn how

other organizations apply Presto

Expert Hadoop 2 Administration

Preserved buildings and historic districts, museums and reconstructions have become an important part of the landscape of cities around the world. Beginning in the 1970s, Tokyo participated in this trend. However, repeated destruction and rapid redevelopment left the city with little building stock of recognized historical value. Late twentieth-century Tokyo thus presents an illuminating case of the emergence of a new sense of history in the city's physical environment, since it required both a shift in perceptions of value and a search for history in the margins and interstices of a rapidly modernizing cityscape. Scholarship to date has tended to view historicism in the postindustrial context as either a genuine response to loss, or as a cynical commodification of the past. The historical process of Tokyo's historicization suggests other interpretations. Moving from the politics of the public square to the invention of neighborhood community, to oddities found and appropriated in the streets, to the consecration of everyday scenes and artifacts as heritage in museums, Tokyo Vernacular traces the rediscovery of the past—sometimes in unlikely forms—in a city with few traditional landmarks. Tokyo's rediscovered past was mobilized as part of a new politics of the everyday after the failure of mass politics in the 1960s. Rather than conceiving the city as national center and claiming public space as national citizens, the post-1960s

generation came to value the local places and things that embodied the vernacular language of the city, and to seek what could be claimed as common property outside the spaces of corporate capitalism and the state.

High Performance Spark

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

Mastering Apache Cassandra

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