Cassandra The Definitive Guide Distributed Data At Web Scale

The book is aimed at intermediate developers with an understanding of core database concepts who want to become a master at implementing Cassandra for their application.

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

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 third edition—updated for Cassandra 4.0—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 nonrelational 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

What value this book provides? This book absolutely provides tremendous value in terms its usefulness. This book takes away the pain associated with learning and mastering Cassandra. All complexity has been digested by the author and simplified for the reader with very useful and practical information that can be absorbed just by glancing through the pages. Years of author's experience and practical knowledge in Cassandra has been gifted to the reader in this book with great diligence and generosity. If you are planning to undergo expensive Cassandra training, think again, spending few hours with this book will change your mind, this book has been written with great care to reduce the learning curve. The aim of this book is multi fold, just to mention a few: Provide confidence to the reader in Cassandra concepts and architecture Provide a flexible, practical framework and context customizable for various situations Provide practical guidance to manage Cassandra platforms of various hues, sizes, shapes Provide real world examples to reduce guess work Provide executable query statements and command line statements at every step Provide practical outcomes to help the reader to gain instance understanding of what to expect Provide insights into making Cassandra environment robust and scalable Provide tricks and tips to implement and maintain seamlessly Provide security and vulnerability mitigation tips and steps Provide best practices to follow for optimal Cassandra use There is no doubt - this book makes the reader very productive Cassandra professional in very short span of time. This book essentially bridges the training gap as the industry is moving fast to take full advantage of what Cassandra can offer to fulfill emerging business needs. This book can be very helpful to Data administrators, Developers, Data modelers/Application Architects, Project Managers and Data Custodians. This book has range of topics, many are listed here: Cassandra concepts and architecture Cassandra Installation and Configuration Cassandra system architecture depicting gossip protocol, replication, consistency, tombstones, hinted handoff, compactions, repairs, memtables, commit log, read and write path functions Cassandra oriented data modelling Cassandra QL (CQL) tutorial Handling of Primary and Partition keys in Cassandra covering No joins, Static columns and TTL aspects Configuring authentication, authorization to access Cassandra in addition, steps to set up node-node and client-node SSL Configuring nodes addition, removal, decommission in single token and vnode setup modes in Cassandra Instructions to add new data center and delink the existing data center from a multi-dc cluster arrangement Cassandra backup and recovery functions with real examples of restoring tables after truncation events Cassandra utilities such as sstabledump, sstablemetadata, sstablesplit, cqlsh and cassandra-stress Troubleshooting methods such as Node down, Read latency and Recover truncated table Upgrading Cassandra to higher versions Additional Cassandra architecture II methods such as Read and Write path, Compactions and Repairs Written in a clear, step-by-step manner, this 400-page course provides an excellent starting point for people that want to get into Apache Cassandra and learn best by doing. A high-quality, project-based, hands-on training courseware book, Apache Cassandra Hands-On Training Level One is designed to be used as the student book for a 2-day introductory level Cassandra course delivered by a Cassandra instructor. Having said that, this book can also be done as a self-paced training course. Recommended prerequisites for this training book are experience with databases, SQL, and programming. This hands-on training book takes people through the basics of working with Cassandra as they learn how to install Cassandra, create a database, create tables, insert, update, and delete data, and create an application, as well as create and modify a multiple-node cluster. Unit 1: Understanding What Cassandra is For Unit 2: Getting Started with the Architecture Unit 3: Installing Cassandra Unit 4: Communicating with Cassandra Unit 5: Creating a Database Unit 6: Creating a Table Unit 7: Inserting Data Unit 8: Modeling Data Unit 9: Creating an Application Unit 10: Updating and Deleting Data Unit 11: Selecting Hardware Unit 12: Adding Nodes to a Cluster Unit 13: Repairing Nodes Unit 14: Removing a Node Unit 15: Monitoring a Cluster Unit 16: Adding a Data Center As virtual machine images are used extensively throughout this hands-on course, including for the creation of a multiple-node Cassandra cluster, any computer used for the exercises in this course needs to be relatively high spec. Specifically, a computer with the following is needed: 64-bit operating system (Mac, Windows, or Linux) 8GB (or more) of RAM 30GB (or more) of free hard drive space Latest version of VMware Player installed and working A way to unzip files Acrobat Reader (or equivalent, for viewing a PDF file) For the full outline, and class files download, see ruthstryker.com/books/achotl1. For a sample unit, see ruthstryker.com/books/achotl1/achotl1_ch06_20140717.pdf (Unit 6) or ruthstryker.com/books/achotl1/achotl1_ch15_20140717.pdf

(Unit 15). For the setup steps, see ruthstryker.com/books/achotl1/achotl1_apC_20140722.pdf (Appendix C). Student comments about the book: "Excellent starter course that has taken me from knowing nothing of Cassandra to feeling confident in setting up and using it." "Level covered in book is just right." "Course material was good. It had a wide range of labs and was very helpful in understanding the agenda." "Course material was well-written and easy to follow." "Excellent introduction into Cassandra filled with hands-on exercises for all topics." "Material covers the basics quite well."

”Eric and Russell were early adopters of Cassandra at SimpleReach. In Practical Cassandra, you benefit from their experience in the trenches administering Cassandra, developing against it, and building one of the first CQL drivers. If you are deploying Cassandra soon, or you inherited a Cassandra cluster to tend, spend some time with the deployment, performance tuning, and maintenance chapters… If you are new to Cassandra, I highly recommend the chapters on data modeling and CQL.” –From the Foreword by Jonathon Ellis, Apache Cassandra Chair Build and Deploy Massively Scalable, Super-fast Data Management Applications with Apache Cassandra Practical Cassandra is the first hands-on developer's guide to building Cassandra systems and applications that deliver breakthrough speed, scalability, reliability, and performance. Fully up to date, it reflects the latest versions of Cassandra–including Cassandra Query Language (CQL), which dramatically lowers the learning curve for Cassandra developers. Pioneering Cassandra developers and Datastax MVPs Russell Bradberry and Eric Lubow walk you through every step of building a real production application that can store enormous amounts of structured, semi-structured, and unstructured data. Drawing on their exceptional expertise, Bradberry and Lubow share practical insights into issues ranging from querying to deployment, management, maintenance, monitoring, and troubleshooting. The authors cover key issues, from architecture to migration, and guide you through crucial decisions about configuration and data modeling. They provide tested sample code, detailed explanations of how Cassandra works ”under the covers,” and new case studies from three cutting-edge users: Ooyala, Hailo, and eBay. Coverage includes Understanding Cassandra's approach, architecture, key concepts, and primary use cases–and why it's so blazingly fast Getting Cassandra up and running on single nodes and large clusters Applying the new design patterns, philosophies, and features that make Cassandra such a powerful data store Leveraging CQL to simplify your transition from SQL-based RDBMSes Deploying and provisioning through the cloud or on bare-metal hardware Choosing the right configuration options for each type of workload Tweaking Cassandra to get maximum performance from your hardware, OS, and JVM Mastering Cassandra's essential tools for maintenance and monitoring Efficiently solving the most common problems with Cassandra deployment, operation, and application development

Learn how to integrate full-stack open source big data architecture and to choose the correct technology—Scala/Spark, Mesos, Akka, Cassandra, and Kafka—in every layer. Big data architecture is becoming a requirement for many different enterprises. So far, however, the focus has largely been on collecting, aggregating, and crunching large data sets in a timely manner. In many cases now, organizations need more than one paradigm to perform efficient analyses. Big Data SMACK explains each of the full-stack technologies and, more importantly, how to best integrate them. It provides detailed coverage of the practical benefits of these technologies and incorporates real-world examples in every situation. This book focuses on the problems and scenarios solved by the architecture, as well as the solutions provided by every technology. It covers the six main concepts of big data architecture and how integrate, replace, and reinforce every layer: The language: Scala The engine: Spark (SQL, MLib, Streaming, GraphX) The container: Mesos, Docker The view: Akka The storage: Cassandra The message broker: Kafka What You Will Learn: Make big data architecture without using complex Greek letter architectures Build a cheap but effective cluster infrastructure Make queries, reports, and graphs that business demands Manage and exploit unstructured and No-SQL data sources Use tools to monitor the performance of your architecture Integrate all technologies and decide which ones replace and which ones reinforce Who This Book Is For: Developers, data architects, and data scientists looking to integrate the most successful big data open stack architecture and to choose the correct technology in every layer

If you’re like most R users, you have deep knowledge and love for statistics. But as your organization continues to collect huge amounts of data, adding tools such as Apache Spark makes a lot of sense. With this practical book, data scientists and professionals working with large-scale data applications will learn how to use Spark from R to tackle big data and big compute problems. Authors Javier Luraschi, Kevin Kuo, and Edgar Ruiz show you how to use R with Spark to solve different data analysis problems. This book covers relevant data science topics, cluster computing, and issues that should interest even the most advanced users. Analyze, explore, transform, and visualize data in Apache Spark with R Create statistical models to extract information and predict outcomes; automate the process in production-ready workflows Perform analysis and modeling across many machines using distributed computing techniques Use large-scale data from multiple sources and different formats with ease from within Spark Learn about alternative modeling frameworks for graph processing, geospatial analysis, and genomics at scale Dive into advanced topics including custom transformations, real-time data processing, and creating custom Spark extensions

Cassandra: the Definitive GuideDistributed Data at Web ScaleO'Reilly Media

Learn how to build a data science technology stack and perform good data science with repeatable methods. You will learn how to turn data lakes into business assets. The data science technology stack demonstrated in Practical Data Science is built from components in general use in the industry. Data scientist Andreas Vermeulen demonstrates in detail how to build and provision a technology stack to yield repeatable results. He shows you how to apply practical methods to extract actionable business knowledge from data lakes consisting of data from a polyglot of data types and dimensions. What You'll Learn Become fluent in the essential concepts and terminology of data science and data engineering Build and use a technology stack that meets industry criteria Master the methods for retrieving actionable business knowledge Coordinate the handling of polyglot data types in a data lake for repeatable results Who This Book Is For Data scientists and data engineers who are required to convert data from a data lake into actionable knowledge for their business, and students who aspire to be data scientists and data engineers

The need to handle increasingly larger data volumes is one factor driving the adoption of a new class of nonrelational “NoSQL” databases. Advocates of NoSQL databases claim they can be used to build systems that are more performant, scale better, and are easier to program. NoSQL Distilled is a concise but thorough introduction to this rapidly emerging technology. Pramod J. Sadalage and Martin Fowler explain how NoSQL databases work and the ways that they may be a superior alternative to a traditional RDBMS. The authors provide a fast-paced guide to the concepts you need to know in order to evaluate whether NoSQL databases are right for your needs and, if so, which technologies you should explore further. The first part of the book concentrates on core concepts, including schemaless data models, aggregates, new distribution models, the CAP theorem, and map-reduce. In the second part, the authors explore architectural and design issues associated with implementing NoSQL. They also present realistic use cases that demonstrate NoSQL databases at work and feature representative examples using Riak, MongoDB, Cassandra, and Neo4j. In addition, by drawing on Pramod Sadalage's pioneering work,

NoSQL Distilled shows how to implement evolutionary design with schema migration: an essential technique for applying NoSQL databases. The book concludes by describing how NoSQL is ushering in a new age of Polyglot Persistence, where multiple data-storage worlds coexist, and architects can choose the technology best optimized for each type of data access.

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 third edition-updated for Cassandra 4.0-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 nonrelational 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.

Data is getting bigger and more complex by the day, and so are your choices in handling it. Explore some of the most cutting-edge databases available - from a traditional relational database to newer NoSQL approaches - and make informed decisions about challenging data storage problems. This is the only comprehensive guide to the world of NoSQL databases, with in-depth practical and conceptual introductions to seven different technologies: Redis, Neo4J, CouchDB, MongoDB, HBase, Postgres, and DynamoDB. This second edition includes a new chapter on DynamoDB and updated content for each chapter. While relational databases such as MySQL remain as relevant as ever, the alternative, NoSQL paradigm has opened up new horizons in performance and scalability and changed the way we approach data-centric problems. This book presents the essential concepts behind each database alongside hands-on examples that make each technology come alive. With each database, tackle a real-world problem that highlights the concepts and features that make it shine. Along the way, explore five database models - relational, key/value, columnar, document, and graph - from the perspective of challenges faced by real applications. Learn how MongoDB and CouchDB are strikingly different, make your applications faster with Redis and more connected with Neo4J, build a cluster of HBase servers using cloud services such as Amazon's Elastic MapReduce, and more. This new edition brings a brand new chapter on DynamoDB, updated code samples and exercises, and a more up-to-date account of each database's feature set. Whether you're a programmer building the next big thing, a data scientist seeking solutions to thorny problems, or a technology enthusiast venturing into new territory, you will find something to inspire you in this book. What You Need: You'll need a *nix shell (Mac OS or Linux preferred, Windows users will need Cygwin), Java 6 (or greater), and Ruby 1.8.7 (or greater). Each chapter will list the downloads required for that database.

WINNER OF THE 2020 NOBEL PRIZE IN PHYSICS The Road to Reality is the most important and ambitious work of science for a generation. It provides nothing less than a comprehensive account of the physical universe and the essentials of its underlying mathematical theory. It assumes no particular specialist knowledge on the part of the reader, so that, for example, the early chapters give us the vital mathematical background to the physical theories explored later in the book. Roger Penrose's purpose is to describe as clearly as possible our present understanding of the universe and to convey a feeling for its deep beauty and philosophical implications, as well as its intricate logical interconnections. The Road to Reality is rarely less than challenging, but the book is leavened by vivid descriptive passages, as well as hundreds of hand-drawn diagrams. In a single work of colossal scope one of the world's greatest scientists has given us a complete and unrivalled guide to the glories of the universe that we all inhabit. 'Roger Penrose is the most important physicist to work in relativity theory except for Einstein. He is one of the very few people I've met in my life who, without reservation, I call a genius' Lee Smolin

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.

Congratulations! You completed the MongoDB application within the given tight timeframe and there is a party to celebrate your application’s release into production. Although people are congratulating you at the celebration, you are feeling some uneasiness inside. To complete the project on time required making a lot of assumptions about the data, such as what terms meant and how calculations are derived. In addition, the poor documentation about the application will be of limited use to the support team, and not investigating all of the inherent rules in the data may eventually lead to poorly-performing structures in the not-so-distant future. Now, what if you had a time machine and could go back and read this book. You would learn that even NoSQL databases like MongoDB require some level of data modeling. Data modeling is the process of learning about the data, and regardless of technology, this process must be performed for a successful application. You would learn the value of conceptual, logical, and physical data modeling and how each stage increases our knowledge of the data and reduces assumptions and poor design decisions. Read this book to learn how to do data modeling for MongoDB applications, and accomplish these five objectives: Understand how data modeling contributes to the process of learning about the data, and is, therefore, a required technique, even when the resulting database is not relational. That is, NoSQL does not mean NoDataModeling! Know how NoSQL databases differ from traditional relational databases, and where MongoDB fits. Explore each MongoDB object and comprehend how each compares to their data modeling and traditional relational database counterparts, and learn the basics of adding, querying, updating, and deleting data in MongoDB. Practice a streamlined, template-driven approach to performing conceptual, logical, and physical data modeling. Recognize that data modeling does not always have to lead to traditional data models! Distinguish top-down from bottom-up development approaches and complete a top-down case study which ties all of the modeling techniques together. This book is written for anyone who is working with, or will be working with MongoDB, including business analysts, data modelers, database administrators, developers, project managers, and data scientists. There are three sections: In Section I, Getting Started, we will reveal the power of data modeling and the tight connections to data models that exist when designing any type of database (Chapter 1), compare NoSQL with traditional relational databases and where MongoDB fits (Chapter 2), explore each MongoDB object and comprehend how each compares to their data modeling and traditional relational database counterparts (Chapter 3), and explain the basics of adding, querying, updating, and deleting data in MongoDB (Chapter 4). In Section II, Levels of Granularity, we cover Conceptual Data Modeling (Chapter 5), Logical Data Modeling (Chapter 6), and Physical Data Modeling (Chapter 7). Notice the “ing” at the end of each of these chapters. We focus on the process of building each of these models, which is where we gain essential business knowledge. In Section III, Case Study, we will explain both top down and bottom up development approaches and go through a top down case study where we start with business requirements and end with the MongoDB database. This case study will tie together all of the techniques in the previous seven chapters. Nike Senior Data Architect Ryan Smith wrote the foreword. Key points are included at the end of each chapter as a way to reinforce concepts. In addition, this book is loaded with hands-on exercises, along with their answers provided in Appendix A. Appendix B contains all of the book’s references and Appendix C contains a glossary of the terms used throughout the text.

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.

Integrating Hadoop leverages the discipline of data integration and applies it to the Hadoop open-source software framework for storing data on clusters of commodity hardware. It is packed with the need-to-know for managers, architects, designers, and developers responsible for populating Hadoop in the enterprise, allowing you to harness big data and do it in such a way that the solution: · Complies with (and even extends) enterprise standards · Integrates seamlessly with the existing information infrastructure · Fills a critical role within enterprise architecture. Integrating Hadoop covers the gamut of the setup, architecture and possibilities for Hadoop in the organization, including: · Supporting an enterprise information strategy · Organizing for a successful Hadoop rollout · Loading and extracting of data in Hadoop · Managing Hadoop data once it’s in the cluster · Utilizing Spark, streaming data, and master data in Hadoop processes - examples are provided to reinforce concepts.

Summary Redis in Action introduces Redis and walks you through examples that demonstrate how to use it effectively. You'll begin by getting Redis set up properly and then exploring the key-value model. Then, you'll dive into real use cases including simple caching, distributed ad targeting, and more. You'll learn how to scale Redis from small jobs to massive datasets. Experienced developers will appreciate chapters on clustering and internal scripting to make Redis easier to use. About the Technology When you need near-real-time access to a fast-moving data stream, key-value stores like Redis are the way to go. Redis expands on the key-value pattern by accepting a wide variety of data types, including hashes, strings, lists, and other structures. It provides lightning-fast operations on in-memory datasets, and also makes it easy to persist to disk on the fly. Plus, it's free and open source. About this book Redis in Action introduces Redis and the key-value model. You'll quickly dive into real use cases including simple caching, distributed ad targeting, and more. You'll learn how to scale Redis from small jobs to massive datasets and discover how to integrate with traditional RDBMS or other NoSQL stores. Experienced developers will appreciate the in-depth chapters on clustering and internal scripting. Written for developers familiar with database concepts. No prior exposure to NoSQL database concepts nor to Redis itself is required. Appropriate for systems administrators comfortable with programming. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. What's Inside Redis from the ground up Preprocessing real-time data Managing in-memory datasets Pub/sub and configuration Persisting to disk About the Author Dr. Josiah L. Carlson is a seasoned database professional and an active contributor to the Redis community. Table of Contents PART 1 GETTING STARTED Getting to know Redis Anatomy of a Redis web application PART 2 CORE CONCEPTS Commands in Redis Keeping data safe and ensuring performance Using Redis for application support Application components in Redis Search-based applications Building a simple social network PART 3 NEXT STEPS Reducing memory use Scaling Redis Scripting Redis with Lua

Working with unbounded and fast-moving data streams has historically been difficult. But with Kafka Streams and ksqlDB, building stream processing applications is easy and fun. This practical guide shows data engineers how to use these tools to build highly scalable stream processing applications for moving, enriching, and transforming large amounts of data in real time. Mitch Seymour, data services engineer at Mailchimp, explains important stream processing concepts against a backdrop of several interesting business problems. You'll learn the strengths of both Kafka Streams and ksqlDB to help you choose the best tool for each unique stream processing project. Non-Java developers will find the ksqlDB path to be an especially gentle introduction to stream processing. Learn the basics of Kafka and the pub/sub communication pattern Build stateless and stateful stream processing applications using Kafka Streams and ksqlDB Perform advanced stateful operations, including windowed joins and aggregations Understand how stateful processing works under the hood Learn about ksqlDB's data integration features, powered by Kafka Connect Work with different types of collections in ksqlDB and perform push and pull queries Deploy your Kafka Streams and ksqlDB applications to production

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

This practical guide explains you to program and understand the power of Apache Cassandra 3.x. You will explore the integration and interaction of Cassandra components, and explore features such as the token allocation algorithm, CQL3, vnodes, lightweight transactions, and data modelling in detail.

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

A beginner's guide to get you up and running with Cassandra, DynamoDB, HBase, InfluxDB, MongoDB, Neo4j, and Redis Key Features Covers the basics of 7 NoSQL databases and how they are used in the enterprises Quick introduction to MongoDB, DynamoDB, Redis, Cassandra, Neo4j, InfluxDB, and HBase Includes effective techniques for database querying and management Book Description This is the golden age of open source NoSQL databases. With enterprises having to work with large amounts of unstructured data and moving away from expensive monolithic architecture, the adoption of NoSQL databases is rapidly increasing. Being familiar with the popular NoSQL databases and knowing how to use them is a must for budding DBAs and developers. This book introduces you to the different types of NoSQL databases and gets you started with seven of the most popular NoSQL databases used by enterprises today. We start off with a brief overview of what NoSQL databases are, followed by an explanation of why and when to use them. The book then covers the seven most popular databases in each of these categories: MongoDB, Amazon DynamoDB, Redis, HBase, Cassandra, InfluxDB, and Neo4j. The book doesn't go into too much detail about each database but teaches you enough to get started with them. By the end of this book, you will have a thorough understanding of the different NoSQL databases and their functionalities, empowering you to select and use the right database

according to your needs. What you will learn Understand how MongoDB provides high-performance, high-availability, and automatic scaling Interact with your Neo4j instances via database queries, Python scripts, and Java application code Get familiar with common querying and programming methods to interact with Redis Study the different types of problems Cassandra can solve Work with HBase components to support common operations such as creating tables and reading/writing data Discover data models and work with CRUD operations using DynamoDB Discover what makes InfluxDB a great choice for working with time-series data Who this book is for If you are a budding DBA or a developer who wants to get started with the fundamentals of NoSQL databases, this book is for you. Relational DBAs who want to get insights into the various offerings of popular NoSQL databases will also find this book to be very useful.

Java SOA Cookbook offers practical solutions and advice to programmers charged with implementing a service-oriented architecture (SOA) in their organization. Instead of providing another conceptual, high-level view of SOA, this cookbook shows you how to make SOA work. It's full of Java and XML code you can insert directly into your applications and recipes you can apply right away. The book focuses primarily on the use of free and open source Java Web Services technologies -- including Java SE 6 and Java EE 5 tools -- but you'll find tips for using commercially available tools as well. Java SOA Cookbook will help you: Construct XML vocabularies and data models appropriate to SOA applications Build real-world web services using the latest Java standards, including JAX-WS 2.1 and JAX-RS 1.0 for RESTful web services Integrate applications from popular service providers using SOAP, POX, and Atom Create service orchestrations with complete coverage of the WS-BPEL (Business Process Execution Language) 2.0 standard Improve the reliability of SOAP-based services with specifications such as WS-Reliable Messaging Deal with governance, interoperability, and quality-of-service issues The recipes in Java SOA Cookbook will equip you with the knowledge you need to approach SOA as an integration challenge, not an obstacle.

Manage the huMONGOus amount of data collected through your web application with MongoDB. This authoritative introduction—written by a core contributor to the project—shows you the many advantages of using document-oriented databases, and demonstrates how this reliable, high-performance system allows for almost infinite horizontal scalability. This updated second edition provides guidance for database developers, advanced configuration for system administrators, and an overview of the concepts and use cases for other people on your project. Ideal for NoSQL newcomers and experienced MongoDB users alike, this guide provides numerous real-world schema design examples. Get started with MongoDB core concepts and vocabulary Perform basic write operations at different levels of safety and speed Create complex queries, with options for limiting, skipping, and sorting results Design an application that works well with MongoDB Aggregate data, including counting, finding distinct values, grouping documents, and using MapReduce Gather and interpret statistics about your collections and databases Set up replica sets and automatic failover in MongoDB Use sharding to scale horizontally, and learn how it impacts applications Delve into monitoring, security and authentication, backup/restore, and other administrative tasks

With this practical book, architects, CTOs, and CIOs will learn a set of patterns for the practice of architecture, including analysis, documentation, and communication. Author Eben Hewitt shows you how to create holistic and thoughtful technology plans, communicate them clearly, lead people toward the vision, and become a great architect or Chief Architect. This book covers each key aspect of architecture comprehensively, including how to incorporate business architecture, information architecture, data architecture, application (software) architecture together to have the best chance for the system’s success. Get a practical set of proven architecture practices focused on shipping great products using architecture Learn how architecture works effectively with development teams, management, and product management teams through the value chain Find updated special coverage on machine learning architecture Get usable templates to start incorporating into your teams immediately Incorporate business architecture, information architecture, data architecture, and application (software) architecture together

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

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

When it comes to choosing, using, and maintaining a database, understanding its internals is essential. But with so many distributed databases and tools available today, it’s often difficult to understand what each one offers and how they differ. With this practical guide, Alex Petrov guides developers through the concepts behind modern database and storage engine internals. Throughout the book, you’ll explore relevant material gleaned from numerous books, papers, blog posts, and the source code of several open source databases. These resources are listed at the end of parts one and two. You’ll discover that the most significant distinctions among many modern databases reside in subsystems that determine how storage is organized and how data is distributed. This book examines: Storage engines: Explore storage classification and taxonomy, and dive into B-Tree-based and immutable Log Structured storage engines, with differences and use-cases for each Storage building blocks: Learn how database files are organized to build efficient storage, using auxiliary data structures such as Page Cache, Buffer Pool and Write-Ahead Log Distributed systems: Learn step-by-step how nodes and processes connect and build complex communication patterns Database clusters: Which consistency models are commonly used by modern databases and how distributed storage systems achieve consistency

Perform fast interactive analytics against different data sources using the Trino high-performance distributed SQL query engine.

With this practical guide, you'll learn how to conduct analytics on data where it lives, whether it's Hive, Cassandra, a relational database, or a proprietary data store. Analysts, software engineers, and production engineers will learn how to manage, use, and even develop with Trino. Initially developed by Facebook, open source Trino is now used by Netflix, Airbnb, LinkedIn, Twitter, Uber, and many other companies. Matt Fuller, Manfred Moser, and Martin Traverso show you how a single Trino query can combine data from multiple sources to allow for analytics across your entire organization. Get started: Explore Trino's use cases and learn about tools that will help you connect to Trino and query data Go deeper: Learn Trino's internal workings, including how to connect to and query data sources with support for SQL statements, operators, functions, and more Put Trino in production: Secure Trino, monitor workloads, tune queries, and connect more applications; learn how other organizations apply Trino
Building distributed applications is difficult enough without having to coordinate the actions that make them work. This practical guide shows how Apache ZooKeeper helps you manage distributed systems, so you can focus mainly on application logic. Even with ZooKeeper, implementing coordination tasks is not trivial, but this book provides good practices to give you a head start, and points out caveats that developers and administrators alike need to watch for along the way. In three separate sections, ZooKeeper contributors Flavio Junqueira and Benjamin Reed introduce the principles of distributed systems, provide ZooKeeper programming techniques, and include the information you need to administer this service. Learn how ZooKeeper solves common coordination tasks Explore the ZooKeeper API’s Java and C implementations and how they differ Use methods to track and react to ZooKeeper state changes Handle failures of the network, application processes, and ZooKeeper itself Learn about ZooKeeper’s trickier aspects dealing with concurrency, ordering, and configuration Use the Curator high-level interface for connection management Become familiar with ZooKeeper internals and administration tools
Technologists who want their ideas heard, understood, and funded are often told to speak the language of business—without really knowing what that is. This book’s toolkit provides architects, product managers, technology managers, and executives with a shared language—in the form of repeatable, practical patterns and templates—to produce great technology strategies. Author Eben Hewitt developed 39 patterns over the course of a decade in his work as CTO, CIO, and chief architect for several global tech companies. With these proven tools, you can define, create, elaborate, refine, and communicate your architecture goals, plans, and approach in a way that executives can readily understand, approve, and execute. This book covers: Architecture and strategy: Adopt a strategic architectural mindset to make a meaningful material impact Creating your strategy: Define the components of your technology strategy using proven patterns Communicating the strategy: Convey your technology strategy in a compelling way to a variety of audiences Bringing it all together: Employ patterns individually or in clusters for specific problems; use the complete framework for a comprehensive strategy
Graph data closes the gap between the way humans and computers view the world. While computers rely on static rows and columns of data, people navigate and reason about life through relationships. This practical guide demonstrates how graph data brings these two approaches together. By working with concepts from graph theory, database schema, distributed systems, and data analysis, you’ll arrive at a unique intersection known as graph thinking. Authors Denise Koessler Gosnell and Matthias Broecheler show data engineers, data scientists, and data analysts how to solve complex problems with graph databases. You’ll explore templates for building with graph technology, along with examples that demonstrate how teams think about graph data within an application. Build an example application architecture with relational and graph technologies Use graph technology to build a Customer 360 application, the most popular graph data pattern today Dive into hierarchical data and troubleshoot a new paradigm that comes from working with graph data Find paths in graph data and learn why your trust in different paths motivates and informs your preferences Use collaborative filtering to design a Netflix-inspired recommendation system
Over 150 recipes to design and optimize large scale Apache Cassandra deployments.
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
Follow this handbook to build, configure, tune, and secure Apache Cassandra databases. Start with the installation of Cassandra and move on to the creation of a single instance, and then a cluster of Cassandra databases. Cassandra is increasingly a key player in many big data environments, and this book shows you how to use Cassandra with Apache Spark, a popular big data processing framework. Also covered are day-to-day topics of importance such as the backup and recovery of Cassandra databases, using the right compression and compaction strategies, and loading and unloading data. Expert Apache Cassandra Administration provides numerous step-by-step examples starting with the basics of a Cassandra database, and going all the way through backup and recovery, performance optimization, and monitoring and securing the data. The book serves as an authoritative and comprehensive guide to the building and management of simple to complex Cassandra databases. The book: Takes you through building a Cassandra database from installation of the software and creation of a single database, through to complex clusters and data centers Provides numerous examples of actual commands in a real-life Cassandra environment that show how to confidently configure, manage, troubleshoot, and tune Cassandra databases Shows how to use the Cassandra configuration properties to build a highly stable, available, and secure Cassandra database that always operates at peak efficiency What You'll Learn Install the Cassandra software and create your first database Understand the Cassandra data model, and the internal architecture of a Cassandra database Create your own Cassandra cluster, step-by-step Run a Cassandra cluster on Docker Work with Apache Spark by connecting to a Cassandra database Deploy Cassandra clusters in your data center, or on Amazon EC2 instances Back up and restore mission-critical Cassandra databases Monitor, troubleshoot, and tune production Cassandra databases, and cut your spending on resources such as memory, servers, and storage Who This Book Is For Database administrators, developers, and architects who are looking for an authoritative and comprehensive single volume for all

their Cassandra administration needs. Also for administrators who are tasked with setting up and maintaining highly reliable and high-performing Cassandra databases. An excellent choice for big data administrators, database administrators, architects, and developers who use Cassandra as their key data store, to support high volume online transactions, or as a decentralized, elastic data store.

Advanced data management has always been at the core of efficient database and information systems. Recent trends like big data and cloud computing have aggravated the need for sophisticated and flexible data storage and processing solutions. This book provides a comprehensive coverage of the principles of data management developed in the last decades with a focus on data structures and query languages. It treats a wealth of different data models and surveys the foundations of structuring, processing, storing and querying data according these models. Starting off with the topic of database design, it further discusses weaknesses of the relational data model, and then proceeds to convey the basics of graph data, tree-structured XML data, key-value pairs and nested, semi-structured JSON data, columnar and record-oriented data as well as object-oriented data. The final chapters round the book off with an analysis of fragmentation, replication and consistency strategies for data management in distributed databases as well as recommendations for handling polyglot persistence in multi-model databases and multi-database architectures. While primarily geared towards students of Master-level courses in Computer Science and related areas, this book may also be of benefit to practitioners looking for a reference book on data modeling and query processing. It provides both theoretical depth and a concise treatment of open source technologies currently on the market.

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 third edition--updated for Cassandra 4.0--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 nonrelational 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

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

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