

Pro Apache Beehive Experts Choice

The professional's one-stop guide to this open-source, Java-based big data framework Professional Hadoop is the complete reference and resource for experienced developers looking to employ Apache Hadoop in real-world settings. Written by an expert team of certified Hadoop developers, committers, and Summit speakers, this book details every key aspect of Hadoop technology to enable optimal processing of large data sets. Designed expressly for the professional developer, this book skips over the basics of database development to get you acquainted with the framework's processes and capabilities right away. The discussion covers each key Hadoop component individually, culminating in a sample application that brings all of the pieces together to illustrate the cooperation and interplay that make Hadoop a major big data solution. Coverage includes everything from storage and security to computing and user experience, with expert guidance on integrating other software and more. Hadoop is quickly reaching significant market usage, and more and more developers are being called upon to develop big data solutions using the Hadoop framework. This book covers the process from beginning to end, providing a crash course for professionals needing to learn and apply Hadoop quickly. Configure storage, UE, and in-memory computing Integrate Hadoop with other programs including Kafka and Storm Master the fundamentals of Apache Big Top and Ignite Build robust data security with expert tips and advice Hadoop's popularity is largely due to its accessibility. Open-source and written in Java, the framework offers almost no barrier to entry for experienced database developers already familiar with the skills and requirements real-world programming entails. Professional Hadoop gives you the practical information and framework-specific skills you need quickly.

Learn the right cutting-edge skills and knowledge to leverage Spark Streaming to implement a wide array of real-time, streaming applications. This book walks you through end-to-end real-time application development using real-world applications, data, and code. Taking an application-first approach, each chapter introduces use cases from a specific industry and uses publicly available datasets from that domain to unravel the intricacies of production-grade design and implementation. The domains covered in Pro Spark Streaming include social media, the sharing economy, finance, online advertising, telecommunication, and IoT. In the last few years, Spark has become synonymous with big data processing. DStreams enhance the underlying Spark processing engine to support streaming analysis with a novel micro-batch processing model. Pro Spark Streaming by Zubair Nabi will enable you to become a specialist of latency sensitive applications by leveraging the key features of DStreams, micro-batch processing, and functional programming. To this end, the book includes ready-to-deploy examples and actual code. Pro Spark Streaming will act as the bible of Spark Streaming. What You'll Learn Discover Spark Streaming application development and best practices Work with the low-level details of discretized streams Optimize production-grade deployments of Spark Streaming via configuration recipes and instrumentation using Graphite, collectd, and Nagios Ingest data from disparate sources including MQTT, Flume, Kafka, Twitter, and a custom HTTP receiver Integrate and couple with HBase, Cassandra, and Redis Take advantage of design patterns for side-effects and maintaining state across the Spark Streaming micro-batch model Implement real-time and scalable ETL using data frames, SparkSQL, Hive, and SparkR Use streaming machine learning, predictive analytics, and recommendations Mesh batch processing with stream processing via the Lambda architecture Who This Book Is For Data scientists, big data experts, BI analysts, and data architects.

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

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

There is an easier way to build Hadoop applications. With this hands-on book, you'll learn how to use Cascading, the open source abstraction framework for Hadoop that lets you easily create and manage powerful enterprise-grade data processing applications—without having to learn the intricacies of MapReduce. Working with sample apps based on Java and other JVM languages, you'll quickly learn Cascading's streamlined approach to data processing, data filtering, and workflow optimization. This book demonstrates how this framework can help your business extract meaningful information from large amounts of distributed data. Start working on Cascading example projects right away Model and analyze unstructured data in any format, from any source Build and test applications with familiar constructs and reusable components Work with the Scalding and Cascalog Domain-Specific Languages Easily deploy applications to Hadoop, regardless of cluster location or data size Build workflows that integrate several big data frameworks and processes Explore common use cases for Cascading, including features and tools that support them Examine a case study that uses a dataset from the Open Data Initiative Could be first to market book on Pro Apache Geronimo Apache Geronimo is open source lightweight (like Spring, Hibernate and Apache Beehive), enterprise Java deployment tool Practical, hands on book

with lots of code samples to learn and apply

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.

This book is a complete practical approach for Hadoop lovers. It is mainly aimed at beginners who want to have a hands-on experience with Hadoop and its ecosystem. Its simplicity and step-by-step explanation will help students and other readers in the computer science industry to use this book as a reference manual. The book has been divided into various chapters that cover Hadoop installation, Summary on Hadoop core components, General commands in Hadoop with examples, SMOOP-import & export commands with verification steps, Pig Latin Commands, Analysis using Pig Latin, Pig Script examples, HiveQL Queries and expected outputs and HBase with CRUD operations. In short, this book is a guide for programmers and non-programmers to begin their projects in Hadoop. It is also suitable as a reference manual for students and professionals who are new to the Hadoop Ecosystems.

Virtual, hands-on learning labs allow you to apply your technical skills in realistic environments. So Sybex has bundled AWS labs from XtremeLabs with our popular AWS Certified Data Analytics Study Guide to give you the same experience working in these labs as you prepare for the Certified Data Analytics Exam that you would face in a real-life application. These labs in addition to the book are a proven way to prepare for the certification and for work as an AWS Data Analyst. AWS Certified Data Analytics Study Guide: Specialty (DAS-C01) Exam is intended for individuals who perform in a data analytics-focused role. This UPDATED exam validates an examinee's comprehensive understanding of using AWS services to design, build, secure, and maintain analytics solutions that provide insight from data. It assesses an examinee's ability to define AWS data analytics services and understand how they integrate with each other; and explain how AWS data analytics services fit in the data lifecycle of collection, storage, processing, and visualization. The book focuses on the following domains: • Collection • Storage and Data Management • Processing • Analysis and Visualization • Data Security This is your opportunity to take the next step in your career by expanding and validating your skills on the AWS cloud. AWS is the frontrunner in cloud computing products and services, and the AWS Certified Data Analytics Study Guide: Specialty exam will get you fully prepared through expert content, and real-world knowledge, key exam essentials, chapter review questions, and much more. Written by an AWS subject-matter expert, this study guide covers exam concepts, and provides key review on exam topics. Readers will also have access to Sybex's superior online interactive learning environment and test bank, including chapter tests, practice exams, a glossary of key terms, and electronic flashcards. And included with this version of the book, XtremeLabs virtual labs that run from your browser. The registration code is included with the book and gives you 6 months of unlimited access to XtremeLabs AWS Certified Data Analytics Labs with 3 unique lab modules based on the book.

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

Until recently, Hadoop deployments existed on hardware owned and run by organizations. Now, of course, you can acquire the computing resources and network connectivity to run Hadoop clusters in the cloud. But there's a lot more to deploying Hadoop to the public cloud than simply renting machines. This hands-on guide shows developers and systems administrators familiar with Hadoop how to install, use, and manage cloud-born clusters efficiently. You'll learn how to architect clusters that work with cloud-provider features—not just to avoid pitfalls, but also to take full advantage of these services. You'll also compare the Amazon, Google, and Microsoft clouds, and learn how to set up clusters in each of them. Learn how Hadoop clusters run in the cloud, the problems they can help you solve, and their potential drawbacks Examine the common concepts of cloud providers, including compute capabilities, networking and security, and storage Build a functional Hadoop cluster on cloud infrastructure, and learn what the major providers require Explore use cases for high availability, relational data with Hive, and complex analytics with Spark Get patterns and practices for running cloud clusters, from designing for price and security to dealing with maintenance

Atlanta magazine's editorial mission is to engage our community through provocative writing, authoritative reporting, and superlative design that illuminate the people, the issues, the trends, and the events that define our city. The magazine informs, challenges, and entertains our readers each month while helping them make intelligent choices, not only about what they do and where they go, but what they think about matters of importance to the community and the region. Atlanta magazine's editorial mission is to engage our community through provocative writing, authoritative reporting, and superlative design that illuminate the people, the issues, the trends, and the events that define our city. The magazine informs, challenges, and entertains our readers each month while helping them make intelligent choices, not only about what they do and where they go, but what they think about matters of importance to the community and the region.

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

Data is bigger, arrives faster, and comes in a variety of formats—and it all needs to be processed at scale for analytics or machine learning. But how can you process such varied workloads efficiently? Enter Apache Spark. Updated to include Spark 3.0, this second edition shows data engineers and data scientists why structure and unification in Spark matters. Specifically, this book explains how to perform simple and complex data analytics and employ machine learning algorithms. Through step-by-step walk-throughs, code snippets, and notebooks, you'll be able to: Learn Python, SQL, Scala, or Java high-level Structured APIs Understand Spark operations and SQL Engine Inspect, tune, and debug Spark operations with Spark configurations and Spark UI Connect to data sources: JSON, Parquet, CSV, Avro, ORC, Hive, S3, or Kafka Perform analytics on batch and streaming data using Structured Streaming Build reliable data pipelines with open source Delta Lake and Spark Develop machine learning pipelines with MLlib and productionize models using MLflow

In this fast-paced book on the Docker open standards platform for developing, packaging and running portable distributed applications, Deepak Vorhadiscusses how to build, ship and run applications on any platform such as a PC, the cloud, data center or a virtual machine. He describes how to install and create Docker images. and the advantages off Docker containers.The remainder of the book is devoted to discussing using Docker with important software solutions. He begins by discussing using Docker with a traditional RDBMS using Oracle and MySQL. Next he moves on to NoSQL with chapter on MongoDB Cassandra, and Couchbase. Then he addresses the use of Docker in the Hadoop ecosystem with complete chapters on utilizing not only Hadoop, but Hive, HBase, Sqoop, Kafka, Solr and Spark. What You Will Learn How to install a Docker image How to create a Docker container How to run an Application in a Docker Container Use Docker with Apache Hadoop Ecosystem Use Docker with NoSQL Databases Use Docker with RDBMS Who This Book Is ForApache Hadoop Developers. Database developers. NoSQL Developers.

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

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

In 2019, America is bordering on financial collapse after engaging in a third conflict in Iraq. On the home front, the president has provided funding for stem cell research, but medical success is overshadowed after corporate greed intervenes. The intention to benefit those with the greatest need quickly disappears as the replication process is used to clone all major organs. With the advent of cloning, the inevitable occurs when Americas population reaches the saturation point. A chip is now implanted into all citizens, which signals when someone reaches the mandatory age of death. Years left on a chip are for sale, and the hunt for new chips is on. Meanwhile, overseas, the war rages, where people die irrespective of age. Lieutenant Roger DeMarco must put an end to the murder of entire innocent Iraqi villages by those who would harvest the enemy and civilians alike. Doctors are caught on opposing sides of science, and even a reality TV star becomes of national interest when he sells his chip and plans to slowly kill himself in a parade of hedonism. Progress, profit, and morality collide with stunning ferocity as America must evolve or die.

Unlock the power of your data with Hadoop 2.X ecosystem and its data warehousing techniques across large data sets About This Book Conquer the mountain of data using Hadoop 2.X tools The authors succeed in creating a context for Hadoop and its ecosystem Hands-on examples and recipes giving the bigger picture and helping you to master Hadoop 2.X data processing platforms Overcome the challenging data processing problems using this exhaustive course with Hadoop 2.X Who This Book Is For This course is for Java developers, who know scripting, wanting a career shift to Hadoop - Big Data segment of the IT industry. So if you are a novice in Hadoop or an expert, this book will make you reach the most advanced level in Hadoop 2.X. What You Will Learn Best practices for setup and configuration of Hadoop clusters, tailoring the system to the problem at hand Integration with relational databases, using Hive for SQL queries and Sqoop for data transfer Installing and maintaining Hadoop 2.X cluster and its ecosystem Advanced Data Analysis using the Hive, Pig, and Map Reduce programs Machine learning principles with libraries such as Mahout and Batch and Stream data processing using Apache Spark Understand the changes involved in the process in the move from Hadoop 1.0 to Hadoop 2.0 Dive into YARN and Storm and use YARN to integrate Storm with Hadoop Deploy Hadoop on Amazon Elastic MapReduce and Discover HDFS replacements and learn about HDFS Federation In Detail As Marc Andreessen has said "Data is eating the world," which can be witnessed today being the age of Big Data, businesses are producing data in huge volumes every day and this rise in tide of data need to be organized and analyzed in a more secured way. With proper and effective use of Hadoop, you can build new-improved models, and based on that you will be able to make the right decisions. The first module, Hadoop beginners Guide will walk you through on understanding Hadoop with very detailed instructions and how to go about using it. Commands are explained using sections called "What just happened" for more clarity and understanding. The second module, Hadoop Real World Solutions Cookbook, 2nd edition, is an essential tutorial to effectively implement a big data warehouse in your business, where you get detailed practices on the latest technologies such as YARN and Spark. Big data has become a key basis of competition and the new waves of productivity growth. Hence, once you get familiar with the basics and implement the end-to-end big data use cases, you will start exploring the third module, Mastering Hadoop. So, now the question is if you need to broaden your Hadoop skill set to the next level after you nail the basics and the advance concepts, then this course is indispensable. When you finish this course, you will be able to tackle the real-world scenarios and become a big data expert using the tools and the knowledge based on the various step-by-step tutorials and recipes. Style and approach This course has covered everything right from the basic concepts of Hadoop till you master the advance mechanisms to become

a big data expert. The goal here is to help you learn the basic essentials using the step-by-step tutorials and from there moving toward the recipes with various real-world solutions for you. It covers all the important aspects of Hadoop from system designing and configuring Hadoop, machine learning principles with various libraries with chapters illustrated with code fragments and schematic diagrams. This is a compendious course to explore Hadoop from the basics to the most advanced techniques available in Hadoop 2.X.

Move your career forward with AWS certification! Prepare for the AWS Certified Data Analytics Specialty Exam with this thorough study guide This comprehensive study guide will help assess your technical skills and prepare for the updated AWS Certified Data Analytics exam. Earning this AWS certification will confirm your expertise in designing and implementing AWS services to derive value from data. The AWS Certified Data Analytics Study Guide: Specialty (DAS-C01) Exam is designed for business analysts and IT professionals who perform complex Big Data analyses. This AWS Specialty Exam guide gets you ready for certification testing with expert content, real-world knowledge, key exam concepts, and topic reviews. Gain confidence by studying the subject areas and working through the practice questions. Big data concepts covered in the guide include: Collection Storage Processing Analysis Visualization Data security AWS certifications allow professionals to demonstrate skills related to leading Amazon Web Services technology. The AWS Certified Data Analytics Specialty (DAS-C01) Exam specifically evaluates your ability to design and maintain Big Data, leverage tools to automate data analysis, and implement AWS Big Data services according to architectural best practices. An exam study guide can help you feel more prepared about taking an AWS certification test and advancing your professional career. In addition to the guide's content, you'll have access to an online learning environment and test bank that offers practice exams, a glossary, and electronic flashcards.

This book provides an introduction to Apache HTTP Server - a free, open-source web server. Apache is the most popular web server software on the Internet; it is estimated that 50% of all active websites use Apache as their web server. You will learn how to download and install Apache HTTP Server on your Windows and Linux system, how to configure Apache as a web server, a proxy server, and a reverse proxy server. You will also learn to set up SSL and password-protect directories on your web server. Later in the book we explain modules and how you can use them to add more features to your web server. The topics covered in this book are: downloading and installing Apache HTTP Server on Ubuntu and Windows understanding Apache configuration files using virtual hosts to hold multiple websites on a single server enabling SSL for secure connections what are modules and how to use them to expand Apache functionality configuring Apache as a forward or reverse proxy redirecting URLs Note that this book uses Ubuntu as an underlying Linux distribution, so some of the commands and configurations files might differ if you are using some other non-Debian based Linux distribution.

* Pro Apache Beehive should be the first and only book on Apache Beehive including its Eclipse Pollinate plug-in at the time it is published. * Covers this much anticipated open source SOA-driven J2EE alternative framework, originally created by BEA and later contributed to Apache. * In-depth, hands-on coverage including Comparison between PageFlows in Workshop to the Standard.

Utilize this practical and easy-to-follow guide to modernize traditional enterprise data warehouse and business intelligence environments with next-generation big data technologies. Next-Generation Big Data takes a holistic approach, covering the most important aspects of modern enterprise big data. The book covers not only the main technology stack but also the next-generation tools and applications used for big data warehousing, data warehouse optimization, real-time and batch data ingestion and processing, real-time data visualization, big data governance, data wrangling, big data cloud deployments, and distributed in-memory big data computing. Finally, the book has an extensive and detailed coverage of big data case studies from Navistar, Cerner, British Telecom, Shopzilla, Thomson Reuters, and Mastercard. What You'll Learn Install Apache Kudu, Impala, and Spark to modernize enterprise data warehouse and business intelligence environments, complete with real-world, easy-to-follow examples, and practical advice Integrate HBase, Solr, Oracle, SQL Server, MySQL, Flume, Kafka, HDFS, and Amazon S3 with Apache Kudu, Impala, and Spark Use StreamSets, Talend, Pentaho, and CDAP for real-time and batch data ingestion and processing Utilize Trifacta, Alteryx, and Datameer for data wrangling and interactive data processing Turbocharge Spark with Alluxio, a distributed in-memory storage platform Deploy big data in the cloud using Cloudera Director Perform real-time data visualization and time series analysis using Zoomdata, Apache Kudu, Impala, and Spark Understand enterprise big data topics such as big data governance, metadata management, data lineage, impact analysis, and policy enforcement, and how to use Cloudera Navigator to perform common data governance tasks Implement big data use cases such as big data warehousing, data warehouse optimization, Internet of Things, real-time data ingestion and analytics, complex event processing, and scalable predictive modeling Study real-world big data case studies from innovative companies, including Navistar, Cerner, British Telecom, Shopzilla, Thomson Reuters, and Mastercard Who This Book Is For BI and big data warehouse professionals interested in gaining practical and real-world insight into next-generation big data processing and analytics using Apache Kudu, Impala, and Spark; and those who want to learn more about other advanced enterprise topics

Beginning Apache Cassandra Development introduces you to one of the most robust and best-performing NoSQL database platforms on the planet. Apache Cassandra is a document database following the JSON document model. It is specifically designed to manage large amounts of data across many commodity servers without there being any single point of failure. This design approach makes Apache Cassandra a robust and easy-to-implement platform when high availability is needed. Apache Cassandra can be used by developers in Java, PHP, Python, and JavaScript—the primary and most commonly used languages. In Beginning Apache Cassandra Development, author and Cassandra expert Vivek Mishra takes you through using Apache Cassandra from each of these primary languages. Mishra also covers the Cassandra Query Language (CQL), the Apache Cassandra analog to SQL. You'll learn to develop applications sourcing data from Cassandra, query that data, and deliver it at speed to your application's users. Cassandra is one of the leading NoSQL databases, meaning you get unparalleled throughput and performance without the sort of processing overhead that comes with traditional proprietary databases. Beginning Apache Cassandra Development will therefore help you create applications that generate search results quickly, stand up to high levels of demand, scale as your user base grows, ensure operational simplicity, and—not least—provide delightful user experiences.

Pro MongoDB Development is about MongoDB, a NoSQL database based on the BSON (binary JSON) document model. The book discusses all aspects of using MongoDB in web applications: Java, PHP, Ruby, JavaScript are the most commonly used programming/scripting languages and the book discusses accessing MongoDB database with these languages. The book also discusses using Java EE frameworks Kundera and Spring Data with MongoDB. As NoSQL databases are commonly used with the Hadoop ecosystem the book also discusses using MongoDB with Apache Hive. Migration from other

NoSQL databases (Apache Cassandra and Couchbase) and from relational databases (Oracle Database) is also discussed. What You'll Learn: How to use a Java client and MongoDB shell How to use MongoDB with PHP, Ruby, and Node.js as well How to migrate Apache Cassandra tables to MongoDB documents; Couchbase to MongoDB; and transferring data between Oracle and MongoDB How to use Kundera, Spring Data, and Spring XD with MongoDB How to load MongoDB data into Oracle Database and integrating MongoDB with Oracle Database in Oracle Data Integrator Audience: The target audience of the book is NoSQL database developers. Target audience includes Java, PHP and Ruby developers. The book is suitable for an intermediate level course in NoSQL database.

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

A hands-on guide to leveraging NoSQL databases NoSQL databases are an efficient and powerful tool for storing and manipulating vast quantities of data. Most NoSQL databases scale well as data grows. In addition, they are often malleable and flexible enough to accommodate semi-structured and sparse data sets. This comprehensive hands-on guide presents fundamental concepts and practical solutions for getting you ready to use NoSQL databases. Expert author Shashank Tiwari begins with a helpful introduction on the subject of NoSQL, explains its characteristics and typical uses, and looks at where it fits in the application stack. Unique insights help you choose which NoSQL solutions are best for solving your specific data storage needs. Professional NoSQL: Demystifies the concepts that relate to NoSQL databases, including column-family oriented stores, key/value databases, and document databases. Delves into installing and configuring a number of NoSQL products and the Hadoop family of products. Explains ways of storing, accessing, and querying data in NoSQL databases through examples that use MongoDB, HBase, Cassandra, Redis, CouchDB, Google App Engine Datastore and more. Looks at architecture and internals. Provides guidelines for optimal usage, performance tuning, and scalable configurations. Presents a number of tools and utilities relating to NoSQL, distributed platforms, and scalable processing, including Hive, Pig, RRDtool, Nagios, and more.

Hadoop in Action teaches readers how to use Hadoop and write MapReduce programs. The intended readers are programmers, architects, and project managers who have to process large amounts of data offline. Hadoop in Action will lead the reader from obtaining a copy of Hadoop to setting it up in a cluster and writing data analytic programs. The book begins by making the basic idea of Hadoop and MapReduce easier to grasp by applying the default Hadoop installation to a few easy-to-follow tasks, such as analyzing changes in word frequency across a body of documents. The book continues through the basic concepts of MapReduce applications developed using Hadoop, including a close look at framework components, use of Hadoop for a variety of data analysis tasks, and numerous examples of Hadoop in action. Hadoop in Action will explain how to use Hadoop and present design patterns and practices of programming MapReduce. MapReduce is a complex idea both conceptually and in its implementation, and Hadoop users are challenged to learn all the knobs and levers for running Hadoop. This book takes you beyond the mechanics of running Hadoop, teaching you to write meaningful programs in a MapReduce framework. This book assumes the reader will have a basic familiarity with Java, as most code examples will be written in Java. Familiarity with basic statistical concepts (e.g. histogram, correlation) will help the reader appreciate the more advanced data processing examples. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book.

This book takes you on a fantastic journey to discover the attributes of big data using Apache Hive. Key Features Grasp the skills needed to write efficient Hive queries to analyze the Big Data Discover how Hive can coexist and work with other tools within the Hadoop ecosystem Uses practical, example-oriented scenarios to cover all the newly released features of Apache Hive 2.3.3 Book Description In this book, we prepare you for your journey into big data by firstly introducing you to backgrounds in the big data domain, alongwith the process of setting up and getting familiar with your Hive working environment. Next, the book guides you through discovering and transforming the values of big data with the help of examples. It also hones your skills in using the Hive language in an efficient manner. Toward the end, the book focuses on advanced topics, such as performance, security, and extensions in Hive, which will guide you on exciting adventures on this worthwhile big data journey. By the end of the book, you will be familiar with Hive and able to work effeciently to find solutions to big data problems What you will learn Create and set up the Hive environment Discover how to use Hive's definition language to describe data Discover interesting data by joining and filtering datasets in Hive Transform data by using Hive sorting, ordering, and functions Aggregate and sample data in different ways Boost Hive query performance and enhance data security in Hive Customize Hive to your needs by using user-defined functions and integrate it with other tools Who this book is for If you are a data analyst, developer, or simply someone who wants to quickly get started with Hive to explore and analyze Big Data in Hadoop, this is the book for you. Since Hive is an SQL-like language, some previous experience with SQL will be useful to get the most out of this book.

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

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

Work with petabyte-scale datasets while building a collaborative, agile workplace in the process. This practical book is the canonical reference to Google BigQuery, the query engine that lets you conduct interactive analysis of large datasets. BigQuery enables enterprises to efficiently store, query, ingest, and learn from their data in a convenient framework. With this book, you'll examine how to analyze data at scale to derive insights from large datasets efficiently. Valliappa Lakshmanan, tech lead for Google Cloud Platform, and Jordan Tigani, engineering director for the BigQuery team, provide best practices for modern data warehousing within an autoscaled, serverless public cloud. Whether you want to explore parts of BigQuery you're not familiar with or prefer to focus on specific tasks, this reference is indispensable.

Pro Microsoft HDInsight is a complete guide to deploying and using Apache Hadoop on the Microsoft Windows Azure Platforms. The information in this book enables you to process enormous volumes of structured as well as non-structured data easily using HDInsight, which is Microsoft's own distribution of Apache Hadoop. Furthermore, the blend of Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) offerings available through Windows Azure lets you take advantage of Hadoop's processing power without the worry of creating, configuring, maintaining, or managing your own cluster. With the data explosion that is soon to happen, the open source Apache Hadoop Framework is gaining traction, and it benefits from a huge ecosystem that has risen around the core functionalities of the Hadoop distributed file system (HDFS™) and Hadoop Map Reduce. Pro Microsoft HDInsight equips you with the knowledge, confidence, and technique to configure and manage this ecosystem on Windows Azure. The book is an excellent choice for anyone aspiring to be a data scientist or data engineer, putting you a step ahead in the data mining field. Guides you through installation and configuration of an HDInsight cluster on Windows Azure Provides clear examples of configuring and executing Map Reduce jobs Helps you consume data and diagnose errors from the Windows Azure HDInsight Service

If you are a data analyst, developer, or simply someone who wants to use Hive to explore and analyze data in Hadoop, this is the book for you. Whether you are new to big data or an expert, with this book, you will be able to master both the basic and the advanced features of Hive. Since Hive is an SQL-like language, some previous experience with the SQL language and databases is useful to have a better understanding of this book.

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

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

[Copyright: 40f2ac021946f79f55d154fed73ff39b](#)