
File Type PDF Pdf Hadoop Mastering

Getting the books **Pdf Hadoop Mastering** now is not type of challenging means. You could not forlorn going like books store or library or borrowing from your links to get into them. This is an definitely easy means to specifically acquire lead by on-line. This online revelation Pdf Hadoop Mastering can be one of the options to accompany you later having supplementary time.

It will not waste your time. take me, the e-book will categorically melody you supplementary thing to read. Just invest tiny times to admission this on-line proclamation **Pdf Hadoop Mastering** as capably as review them wherever you are now.

KEY=MASTERING - CAITLYN BLACKBURN

Mastering Hadoop 3 Big data processing at scale to unlock unique business insights Packt Publishing Ltd A comprehensive guide to mastering the most advanced Hadoop 3 concepts Key Features Get to grips with the newly introduced features and capabilities of Hadoop 3 Crunch and process data using MapReduce, YARN, and a host of tools within the Hadoop ecosystem Sharpen your Hadoop skills with real-world case studies and code Book Description Apache Hadoop is one of the most popular big data solutions for distributed storage and for processing large chunks of data. With Hadoop 3, Apache promises to provide a high-performance, more fault-tolerant, and highly efficient big data processing platform, with a focus on improved scalability and increased efficiency. With this guide, you'll understand advanced concepts of the Hadoop ecosystem tool. You'll learn how Hadoop works internally, study advanced concepts of different ecosystem tools, discover solutions to real-world use cases, and understand how to secure your cluster. It will then walk you through HDFS, YARN, MapReduce, and Hadoop 3 concepts. You'll be able to address common challenges like using Kafka efficiently, designing low latency, reliable message delivery Kafka systems, and handling high data volumes. As you advance, you'll discover how to address major challenges when building an enterprise-grade messaging system, and how to use different stream processing systems along with Kafka to fulfil your enterprise goals. By the end of this book, you'll have a complete understanding of how components in the Hadoop ecosystem are effectively integrated to implement a fast and reliable data pipeline, and you'll be equipped to tackle a range of real-world problems in data pipelines. What you will learn Gain an in-depth understanding of distributed computing using Hadoop 3 Develop enterprise-grade applications using Apache Spark, Flink, and more Build scalable and high-performance Hadoop data pipelines with security, monitoring, and data governance Explore batch data processing patterns and how to model data in Hadoop Master best practices for enterprises using, or planning to use, Hadoop 3 as a data platform Understand security aspects of Hadoop, including authorization and authentication Who this book is for If you want to become a big data professional by mastering the advanced concepts of Hadoop, this book is for you. You'll also find this book useful if you're a Hadoop professional looking to strengthen your knowledge of the Hadoop ecosystem. Fundamental knowledge of the Java programming language and basics of Hadoop is necessary to get started with this book. **Hadoop in Practice Simon and Schuster** 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 **Hadoop For Dummies John Wiley & Sons** Let Hadoop For Dummies help harness the power of your data and rein in the information overload Big data has become big business, and companies and organizations of all sizes are struggling to find ways to retrieve valuable information from their massive data sets with becoming overwhelmed. Enter Hadoop and this easy-to-understand For Dummies guide. Hadoop For Dummies helps readers understand the value of big data, make a business case for using Hadoop, navigate the Hadoop ecosystem, and build and manage Hadoop applications and clusters. Explains the origins of Hadoop, its economic benefits, and its functionality and practical applications Helps you find your way around the Hadoop ecosystem, program MapReduce, utilize design patterns, and get your Hadoop cluster up and running quickly and easily Details how to use Hadoop applications for data mining, web analytics and personalization, large-scale text processing, data science, and problem-solving Shows you how to improve the value of your Hadoop cluster, maximize your investment in Hadoop, and avoid common pitfalls when building your Hadoop cluster From programmers challenged with building and maintaining affordable, scaleable data systems to administrators who must deal with huge volumes of information effectively and efficiently, this how-to has something to help you with Hadoop. **Mastering Hadoop 3 Big Data Processing at Scale to Unlock Unique Business Insights** A comprehensive guide to mastering the most advanced Hadoop 3 concepts Key Features Get to grips with the newly introduced features and capabilities of Hadoop 3 Crunch and process data using MapReduce, YARN, and a host of tools within the Hadoop ecosystem Sharpen your Hadoop skills with real-world case studies and code Book Description Apache Hadoop is one of the most popular big data solutions for distributed storage and for processing large chunks of data. With Hadoop 3, Apache promises to provide a high-performance, more fault-tolerant, and highly efficient big data processing platform, with a focus on improved scalability and increased efficiency. With this guide, you'll understand advanced concepts of the Hadoop ecosystem tool. You'll learn how Hadoop works internally, study advanced concepts of different ecosystem tools, discover solutions to real-world use cases, and understand how to secure your cluster. It will then walk you through HDFS, YARN, MapReduce, and Hadoop 3 concepts. You'll be able to address common challenges like using Kafka efficiently, designing low latency, reliable message delivery Kafka systems, and handling high data volumes. As you advance, you'll discover how to address major challenges when building an enterprise-grade messaging system, and how to use different stream processing systems along with Kafka to fulfil your enterprise goals. By the end of this book, you'll have a complete understanding of how components in the Hadoop ecosystem are effectively integrated to implement a fast and reliable data pipeline, and you'll be equipped to tackle a range of real-world problems in data pipelines. What you will learn Gain an in-depth understanding of distributed computing using Hadoop 3 Develop enterprise-grade applications using Apache Spark, Flink, and more Build scalable and high-performance Hadoop data pipelines with security, monitoring, and data governance Explore batch data processing patterns and how to model data in Hadoop Master best practices for enterprises using, or planning to use, Hadoop 3 as a data platform Understand security aspects of Hadoop, including authorization and authentication Who this book is for If you want to become a big data professional by mastering the advanced concepts of Hadoop, this book is for you. You'll also find this book useful if you're a Hadoop professional looking to strengthen your knowledge of the Hadoop ecosystem. Fundamental knowledge of the Java programming language and basics of Hadoop is necessary to get started with this book. **Hadoop Essentials Packt Publishing Ltd** If you are a system or application developer interested in learning how to solve practical problems using the Hadoop framework, then this book is ideal for you. This book is also meant for Hadoop professionals who want to find solutions to the different challenges they come across in their Hadoop projects. **Mastering Hadoop Packt Publishing Ltd** Do you want to broaden your Hadoop skill set and take your knowledge to the next level? Do you wish to enhance your knowledge of Hadoop to solve challenging data processing problems? Are your Hadoop jobs, Pig scripts, or Hive queries not working as fast as you intend? Are you looking to understand the benefits of upgrading Hadoop? If the answer is yes to any of these, this book is for you. It assumes novice-level familiarity with Hadoop. **Mastering Apache Storm Packt Publishing Ltd** Master the intricacies of Apache Storm and develop real-time stream processing applications with ease About This Book Exploit the various real-time processing functionalities offered by Apache Storm such as parallelism, data partitioning, and more Integrate Storm with other Big Data technologies like Hadoop, HBase, and Apache Kafka An easy-to-understand guide to effortlessly create distributed applications with Storm Who This Book Is For If you are a Java developer who wants to enter into the world of real-time stream processing applications using Apache Storm, then this book is for you. No previous experience in Storm is required as this book starts from the basics. After finishing this book, you will be able to develop not-so-complex Storm applications. What You Will Learn Understand the core concepts of Apache Storm and real-time processing Follow the steps to deploy multiple nodes of Storm Cluster Create Trident topologies to support various message-processing semantics Make your cluster sharing effective using Storm scheduling Integrate Apache Storm with other Big Data technologies such as Hadoop, HBase, Kafka, and more Monitor the health of your Storm cluster In Detail Apache Storm is a real-time Big Data processing framework that processes large amounts of data reliably, guaranteeing that every message will be processed. Storm allows you to scale your data as it grows, making it an excellent platform to solve your big data problems. This extensive guide will help you understand right from the basics to the advanced topics of Storm. The book begins with a detailed introduction to real-time processing and where Storm fits in to solve these problems. You'll get an understanding of deploying Storm on clusters by writing a basic Storm Hello World example. Next we'll introduce you to Trident and you'll get a clear understanding of how you can develop and deploy a trident topology. We cover topics such as monitoring, Storm Parallelism, scheduler and log processing, in a very easy to understand manner. You will also learn how to integrate Storm with other well-known Big Data technologies such as HBase, Redis, Kafka, and Hadoop to realize the full potential of Storm. With real-world examples and clear explanations, this book will ensure you will have a thorough mastery of Apache Storm. You will be able to use this knowledge to develop efficient, distributed real-time applications to cater to your business needs. Style and approach This easy-to-follow guide is full of examples and real-world applications to help you get an in-depth understanding of Apache Storm. This book covers the basics thoroughly and also delves into the intermediate and slightly advanced concepts of application development with Apache Storm. **Mastering Big Data using Spring Cloud and Hadoop Clusters Dhiraj Baraik** What is Big Data? Big Data is a collection of large datasets that cannot be processed using traditional computing techniques. It is not a single technique or a tool, rather it involves many areas of business and technology. What Comes Under Big Data? Big data involves the data produced by different devices and applications. Given below are some of the fields that come under the umbrella of Big Data. · Black Box Data: It is a component of helicopter, airplanes, and jets, etc. It captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft. · Social Media Data: Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe. · Stock Exchange Data: The stock exchange data holds information about the 'buy' and 'sell' decisions made on a share of different companies made by the customers. · Power Grid Data: The power grid data holds information consumed by a particular node with respect to a base station. · Transport Data: Transport data includes model, capacity, distance and availability of a vehicle. · Search Engine Data: Search engines retrieve lots of data from different databases. **Educational Tool for Hadoop** You do not see a lot of good informative tutorials/information on Hadoop EcoSystem which helps you in mastering Hadoop from beginning to the end. Most of the Hadoop tutorials available on the web describe everything as if there will be no errors given by the system during the learning process and all information about Hadoop is always conveyed in an expert language and not in a simple easy to understand language which at times is difficult

to grasp. The above problem gave rise to have a tool that will provide you a useful learning experience. Dr. Eckberg (Chair of Thesis) expressed his desire to have a tool that will help the Computer Science students of our university to learn Hadoop by just using this tool and with that marked the birth of this tool, called an Educational Tool for Hadoop. The beauty of this tool is that all the tutorials are in the form of pdf which can be opened through the tool and saved for further use at any point of time. The second advantage the tool provides is the simplicity of the language used in these tutorials because these tutorials have been practically designed and performed by the author and then described as simply as possible. The advantage of that it has been made with a typical student mindset in view, and has been created by a master's student very familiar with that mindset, so that the experience that it gives to the end user is hopefully exactly what he needs. This tool covers three essential parts of the Hadoop Ecosystem namely Hive, HBase and Sqoop. The tool will take you through the complete learning process for the above mentioned components. All the practical examples and the outputs attached in the pdf documents that are accessed through this tool make learning even easier. As all tutorials have been designed and performed by the author on a standard laptop and as all the outputs have been attached in the pdf's we can say that the tool fully illustrates the use of Hadoop along with Hive, HBase and Sqoop. **Mastering Spark with R The Complete Guide to Large-Scale Analysis and Modeling "O'Reilly Media, Inc."** 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 **Hadoop: The Definitive Guide "O'Reilly Media, Inc."** Ready to unlock the power of your data? With this comprehensive guide, you'll learn how to build and maintain reliable, scalable, distributed systems with Apache Hadoop. This book is ideal for programmers looking to analyze datasets of any size, and for administrators who want to set up and run Hadoop clusters. You'll find illuminating case studies that demonstrate how Hadoop is used to solve specific problems. This third edition covers recent changes to Hadoop, including material on the new MapReduce API, as well as MapReduce 2 and its more flexible execution model (YARN). Store large datasets with the Hadoop Distributed File System (HDFS) Run distributed computations with MapReduce Use Hadoop's data and I/O building blocks for compression, data integrity, serialization (including Avro), and persistence Discover common pitfalls and advanced features for writing real-world MapReduce programs Design, build, and administer a dedicated Hadoop cluster—or run Hadoop in the cloud Load data from relational databases into HDFS, using Sqoop Perform large-scale data processing with the Pig query language Analyze datasets with Hive, Hadoop's data warehousing system Take advantage of HBase for structured and semi-structured data, and ZooKeeper for building distributed systems **Learning Spark O'Reilly Media** 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 **Mastering Large Datasets Parallelize and Distribute Your Python Code Manning Publications** With an emphasis on clarity, style, and performance, author J.T. Wolohan expertly guides you through implementing a functionally-influenced approach to Python coding. You'll get familiar with Python's functional built-ins like the `functools` operator and `itertools` modules, as well as the `toolz` library. *Mastering Large Datasets* teaches you to write easily readable, easily scalable Python code that can efficiently process large volumes of structured and unstructured data. By the end of this comprehensive guide, you'll have a solid grasp on the tools and methods that will take your code beyond the laptop and your data science career to the next level! Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. **Mastering Large Datasets with Python Parallelize and Distribute Your Python Code Simon and Schuster** Summary Modern data science solutions need to be clean, easy to read, and scalable. In *Mastering Large Datasets with Python*, author J.T. Wolohan teaches you how to take a small project and scale it up using a functionally influenced approach to Python coding. You'll explore methods and built-in Python tools that lend themselves to clarity and scalability, like the high-performing parallelism method, as well as distributed technologies that allow for high data throughput. The abundant hands-on exercises in this practical tutorial will lock in these essential skills for any large-scale data science project. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Programming techniques that work well on laptop-sized data can slow to a crawl—or fail altogether—when applied to massive files or distributed datasets. By mastering the powerful map and reduce paradigm, along with the Python-based tools that support it, you can write data-centric applications that scale efficiently without requiring codebase rewrites as your requirements change. About the book *Mastering Large Datasets with Python* teaches you to write code that can handle datasets of any size. You'll start with laptop-sized datasets that teach you to parallelize data analysis by breaking large tasks into smaller ones that can run simultaneously. You'll then scale those same programs to industrial-sized datasets on a cluster of cloud servers. With the map and reduce paradigm firmly in place, you'll explore tools like Hadoop and PySpark to efficiently process massive distributed datasets, speed up decision-making with machine learning, and simplify your data storage with AWS S3. What's inside An introduction to the map and reduce paradigm Parallelization with the multiprocessing module and pathos framework Hadoop and Spark for distributed computing Running AWS jobs to process large datasets About the reader For Python programmers who need to work faster with more data. About the author J. T. Wolohan is a lead data scientist at Booz Allen Hamilton, and a PhD researcher at Indiana University, Bloomington. Table of Contents: PART 1 1 | Introduction 2 | Accelerating large dataset work: Map and parallel computing 3 | Function pipelines for mapping complex transformations 4 | Processing large datasets with lazy workflows 5 | Accumulation operations with reduce 6 | Speeding up map and reduce with advanced parallelization PART 2 7 | Processing truly big datasets with Hadoop and Spark 8 | Best practices for large data with Apache Streaming and mrjob 9 | PageRank with map and reduce in PySpark 10 | Faster decision-making with machine learning and PySpark PART 3 11 | Large datasets in the cloud with Amazon Web Services and S3 12 | MapReduce in the cloud with Amazon's Elastic MapReduce **Mastering MongoDB 6.x Expert techniques to run high-volume and fault-tolerant database solutions using MongoDB 6.x Packt Publishing Ltd** Design and build solutions with the most powerful document database, MongoDB Key Features Learn from the experts about every new feature in MongoDB 6 and 5 Develop applications and administer clusters using MongoDB on premise or in the cloud Explore code-rich case studies showcasing MongoDB's major features followed by best practices Book Description MongoDB is a leading non-relational database. This book covers all the major features of MongoDB including the latest version 6. MongoDB 6.x adds many new features and expands on existing ones such as aggregation, indexing, replication, sharding and MongoDB Atlas tools. Some of the MongoDB Atlas tools that you will master include Atlas dedicated clusters and Serverless, Atlas Search, Charts, Realm Application Services/Sync, Compass, Cloud Manager and Data Lake. By getting hands-on working with code using realistic use cases, you will master the art of modeling, shaping and querying your data and become the MongoDB oracle for the business. You will focus on broadly used and niche areas such as optimizing queries, configuring large-scale clusters, configuring your cluster for high performance and availability and many more. Later, you will become proficient in auditing, monitoring, and securing your clusters using a structured and organized approach. By the end of this book, you will have grasped all the practical understanding needed to design, develop, administer and scale MongoDB-based database applications both on premises and on the cloud. What you will learn Understand data modeling and schema design, including smart indexing Master querying data using aggregation Use distributed transactions, replication and sharding for better results Administer your database using backups and monitoring tools Secure your cluster with the best checklists and advice Master MongoDB Atlas, Search, Charts, Serverless, Realm, Compass, Cloud Manager and other tools offered in the cloud or on premises Integrate MongoDB with other big data sources Design and deploy MongoDB in mobile, IoT and serverless environments Who this book is for This book is for MongoDB developers and database administrators who want to learn how to model their data using MongoDB in depth, for both greenfield and existing projects. An understanding of MongoDB, shell command skills and basic database design concepts is required to get the most out of this book. **Data Processing and Modeling with Hadoop Mastering Hadoop Ecosystem Including ETL, Data Vault, DMBok, GDPR, and Various Data-Centric Tools (English Edition) BPB Publications** Understand data in a simple way using a data lake. KEY FEATURES ● In-depth practical demonstration of Hadoop/Yarn concepts with numerous examples. ● Includes graphical illustrations and visual explanations for Hadoop commands and parameters. ● Includes details of dimensional modeling and Data Vault modeling. ● Includes details of how to create and define a structure to a data lake. DESCRIPTION The book 'Data Processing and Modeling with Hadoop' explains how a distributed system works and its benefits in the big data era in a straightforward and clear manner. After reading the book, you will be able to plan and organize projects involving a massive amount of data. The book describes the standards and technologies that aid in data management and compares them to other technology business standards. The reader receives practical guidance on how to segregate and separate data into zones, as well as how to develop a model that can aid in data evolution. It discusses security and the measures that are utilized to reduce the impact of security. Self-service analytics, Data Lake, Data Vault 2.0, and Data Mesh are discussed in the book. After reading this book, the reader will have a thorough understanding of how to structure a data lake, as well as the ability to plan, organize, and carry out the implementation of a data-driven business with full governance and security. WHAT YOU WILL LEARN ● Learn the basics of components to the Hadoop Ecosystem. ● Understand the structure, files, and zones of a Data Lake. ● Learn to implement the security part of the Hadoop Ecosystem. ● Learn to work with the Data Vault 2.0 modeling. ● Learn to develop a strategy to define good governance. ● Learn new tools to work with Data and Big Data WHO THIS BOOK IS FOR This book caters to big data developers, technical specialists, consultants, and students who want to build good proficiency in big data. Knowing basic SQL concepts, modeling, and development would be good, although not mandatory. TABLE OF CONTENTS 1. Understanding the Current Moment 2. Defining the Zones 3. The Importance of Modeling 4. Massive Parallel Processing 5. Doing ETL/ELT 6. A Little Governance 7. Talking About Security 8. What Are the Next Steps? **Spark: The Definitive Guide Big Data Processing Made Simple "O'Reilly Media, Inc."** 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 **Data Analytics with Hadoop An Introduction for Data Scientists "O'Reilly Media, Inc."** Ready to use statistical and machine-learning techniques across large data sets? This practical guide shows you why the Hadoop ecosystem is perfect for the job. Instead of deployment, operations, or software development usually associated with distributed computing, you'll focus on particular analyses you can build, the data warehousing techniques that Hadoop provides, and higher order data workflows this framework can produce. Data scientists and analysts will learn how to perform a wide range of techniques, from writing MapReduce and Spark applications with Python to using advanced modeling and data management with Spark MLlib, Hive, and HBase. You'll also learn about the analytical processes and data systems available to build and empower data products that can handle—and actually require—huge amounts of data. Understand core concepts behind Hadoop and cluster computing Use design patterns and parallel analytical algorithms to create distributed data analysis jobs Learn about data management, mining, and warehousing in a distributed context using Apache Hive and HBase Use Sqoop and Apache Flume to ingest data from relational databases Program complex Hadoop and Spark applications with Apache Pig and Spark DataFrames Perform machine learning techniques such as classification, clustering, and collaborative filtering with Spark's MLlib **Mastering MongoDB 3.x An expert's guide to building fault-tolerant MongoDB applications Packt Publishing Ltd** An expert's guide to build fault tolerant MongoDB application About This Book Master the advanced

modeling, querying, and administration techniques in MongoDB and become a MongoDB expert Covers the latest updates and Big Data features frequently used by professional MongoDB developers and administrators If your goal is to become a certified MongoDB professional, this book is your perfect companion Who This Book Is For Mastering MongoDB is a book for database developers, architects, and administrators who want to learn how to use MongoDB more effectively and productively. If you have experience in, and are interested in working with, NoSQL databases to build apps and websites, then this book is for you. What You Will Learn Get hands-on with advanced querying techniques such as indexing, expressions, arrays, and more. Configure, monitor, and maintain highly scalable MongoDB environment like an expert. Master replication and data sharding to optimize read/write performance. Design secure and robust applications based on MongoDB. Administer MongoDB-based applications on-premise or in the cloud Scale MongoDB to achieve your design goals Integrate MongoDB with big data sources to process huge amounts of data In Detail MongoDB has grown to become the de facto NoSQL database with millions of users—from small startups to Fortune 500 companies. Addressing the limitations of SQL schema-based databases, MongoDB pioneered a shift of focus for DevOps and offered sharding and replication maintainable by DevOps teams. The book is based on MongoDB 3.x and covers topics ranging from database querying using the shell, built in drivers, and popular ODM mappers to more advanced topics such as sharding, high availability, and integration with big data sources. You will get an overview of MongoDB and how to play to its strengths, with relevant use cases. After that, you will learn how to query MongoDB effectively and make use of indexes as much as possible. The next part deals with the administration of MongoDB installations on-premise or in the cloud. We deal with database internals in the next section, explaining storage systems and how they can affect performance. The last section of this book deals with replication and MongoDB scaling, along with integration with heterogeneous data sources. By the end this book, you will be equipped with all the required industry skills and knowledge to become a certified MongoDB developer and administrator. Style and approach This book takes a practical, step-by-step approach to explain the concepts of MongoDB. Practical use-cases involving real-world examples are used throughout the book to clearly explain theoretical concepts. **Mastering Apache Spark** Gain expertise in processing and storing data by using advanced techniques with Apache Spark About This Book- Explore the integration of Apache Spark with third party applications such as H2O, Databricks and Titan- Evaluate how Cassandra and Hbase can be used for storage- An advanced guide with a combination of instructions and practical examples to extend the most up-to date Spark functionalities Who This Book Is For If you are a developer with some experience with Spark and want to strengthen your knowledge of how to get around in the world of Spark, then this book is ideal for you. Basic knowledge of Linux, Hadoop and Spark is assumed. Reasonable knowledge of Scala is expected. What You Will Learn- Extend the tools available for processing and storage- Examine clustering and classification using MLlib- Discover Spark stream processing via Flume, HDFS- Create a schema in Spark SQL, and learn how a Spark schema can be populated with data- Study Spark based graph processing using Spark GraphX- Combine Spark with H2O and deep learning and learn why it is useful- Evaluate how graph storage works with Apache Spark, Titan, HBase and Cassandra- Use Apache Spark in the cloud with Databricks and AWS In Detail Apache Spark is an in-memory cluster based parallel processing system that provides a wide range of functionality like graph processing, machine learning, stream processing and SQL. It operates at unprecedented speeds, is easy to use and offers a rich set of data transformations. This book aims to take your limited knowledge of Spark to the next level by teaching you how to expand Spark functionality. The book commences with an overview of the Spark eco-system. You will learn how to use MLlib to create a fully working neural net for handwriting recognition. You will then discover how stream processing can be tuned for optimal performance and to ensure parallel processing. The book extends to show how to incorporate H2O for machine learning, Titan for graph based storage, Databricks for cloud-based Spark. Intermediate Scala based code examples are provided for Apache Spark module processing in a CentOS Linux and Databricks cloud environment. Style and approach This book is an extensive guide to Apache Spark modules and tools and shows how Spark's functionality can be extended for real-time processing and storage with worked examples. **Mastering Machine Learning with Spark 2.x** Packt Publishing Ltd Unlock the complexities of machine learning algorithms in Spark to generate useful data insights through this data analysis tutorial About This Book Process and analyze big data in a distributed and scalable way Write sophisticated Spark pipelines that incorporate elaborate extraction Build and use regression models to predict flight delays Who This Book Is For Are you a developer with a background in machine learning and statistics who is feeling limited by the current slow and "small data" machine learning tools? Then this is the book for you! In this book, you will create scalable machine learning applications to power a modern data-driven business using Spark. We assume that you already know the machine learning concepts and algorithms and have Spark up and running (whether on a cluster or locally) and have a basic knowledge of the various libraries contained in Spark. What You Will Learn Use Spark streams to cluster tweets online Run the PageRank algorithm to compute user influence Perform complex manipulation of DataFrames using Spark Define Spark pipelines to compose individual data transformations Utilize generated models for off-line/on-line prediction Transfer the learning from an ensemble to a simpler Neural Network Understand basic graph properties and important graph operations Use GraphFrames, an extension of DataFrames to graphs, to study graphs using an elegant query language Use K-means algorithm to cluster movie reviews dataset In Detail The purpose of machine learning is to build systems that learn from data. Being able to understand trends and patterns in complex data is critical to success; it is one of the key strategies to unlock growth in the challenging contemporary marketplace today. With the meteoric rise of machine learning, developers are now keen on finding out how can they make their Spark applications smarter. This book gives you access to transform data into actionable knowledge. The book commences by defining machine learning primitives by the MLlib and H2O libraries. You will learn how to use Binary classification to detect the Higgs Boson particle in the huge amount of data produced by CERN particle collider and classify daily health activities using ensemble Methods for Multi-Class Classification. Next, you will solve a typical regression problem involving flight delay predictions and write sophisticated Spark pipelines. You will analyze Twitter data with help of the doc2vec algorithm and K-means clustering. Finally, you will build different pattern mining models using MLlib, perform complex manipulation of DataFrames using Spark and Spark SQL, and deploy your app in a Spark streaming environment. Style and approach This book takes a practical approach to help you get to grips with using Spark for analytics and to implement machine learning algorithms. We'll teach you about advanced applications of machine learning through illustrative examples. These examples will equip you to harness the potential of machine learning, through Spark, in a variety of enterprise-grade systems. **Programming Hive "O'Reilly Media, Inc."** Describes the features and functions of Apache Hive, the data infrastructure for Hadoop. **R for Programmers Mastering the Tools** CRC Press Unlike other books about R, written from the perspective of statistics, R for Programmers: Mastering the Tools is written from the perspective of programmers, providing a channel for programmers with expertise in other programming languages to quickly understand R. The contents are divided into four sections: The first section consists of the basic **Mastering RStudio - Develop, Communicate, and Collaborate with R** Packt Publishing Ltd Harness the power of RStudio to create web applications, R packages, markdown reports and pretty data visualizations About This Book Discover the multi-functional use of RStudio to support your daily work with R code Learn to create stunning, meaningful, and interactive graphs and learn to embed them into easy communicable reports using multiple R packages Develop your own R packages and Shiny web apps to share your knowledge and collaborate with others. Who This Book Is For This book is aimed at R developers and analysts who wish to do R statistical development while taking advantage of RStudio's functionality to ease their development efforts. R programming experience is assumed as well as being comfortable with R's basic structures and a number of functions. What You Will Learn Discover the RStudio IDE and details about the user interface Communicate your insights with R Markdown in static and interactive ways Learn how to use different graphic systems to visualize your data Build interactive web applications with the Shiny framework to present and share your results Understand the process of package development and assemble your own R packages Easily collaborate with other people on your projects by using Git and GitHub Manage the R environment for your organization with RStudio and Shiny server Apply your obtained knowledge about RStudio and R development to create a real-world dashboard solution In Detail RStudio helps you to manage small to large projects by giving you a multi-functional integrated development environment, combined with the power and flexibility of the R programming language, which is becoming the bridge language of data science for developers and analyst worldwide. Mastering the use of RStudio will help you to solve real-world data problems. This book begins by guiding you through the installation of RStudio and explaining the user interface step by step. From there, the next logical step is to use this knowledge to improve your data analysis workflow. We will do this by building up our toolbox to create interactive reports and graphs or even web applications with Shiny. To collaborate with others, we will explore how to use Git and GitHub with RStudio and how to build your own packages to ensure top quality results. Finally, we put it all together in an interactive dashboard written with R. Style and approach An easy-to-follow guide full of hands-on examples to master RStudio. Beginning from explaining the basics, each topic is explained with a lot of details for every feature. **Mastering Apache Spark 2.x** Packt Publishing Ltd Advanced analytics on your Big Data with latest Apache Spark 2.x About This Book An advanced guide with a combination of instructions and practical examples to extend the most up-to date Spark functionalities. Extend your data processing capabilities to process huge chunk of data in minimum time using advanced concepts in Spark. Master the art of real-time processing with the help of Apache Spark 2.x Who This Book Is For If you are a developer with some experience with Spark and want to strengthen your knowledge of how to get around in the world of Spark, then this book is ideal for you. Basic knowledge of Linux, Hadoop and Spark is assumed. Reasonable knowledge of Scala is expected. What You Will Learn Examine Advanced Machine Learning and DeepLearning with MLlib, SparkML, SystemML, H2O and DeepLearning4j Study highly optimised unified batch and real-time data processing using SparkSQL and Structured Streaming Evaluate large-scale Graph Processing and Analysis using GraphX and GraphFrames Apply Apache Spark in Elastic deployments using Jupyter and Zeppelin Notebooks, Docker, Kubernetes and the IBM Cloud Understand internal details of cost based optimizers used in Catalyst, SystemML and GraphFrames Learn how specific parameter settings affect overall performance of an Apache Spark cluster Leverage Scala, R and python for your data science projects In Detail Apache Spark is an in-memory cluster-based parallel processing system that provides a wide range of functionalities such as graph processing, machine learning, stream processing, and SQL. This book aims to take your knowledge of Spark to the next level by teaching you how to expand Spark's functionality and implement your data flows and machine/deep learning programs on top of the platform. The book commences with an overview of the Spark ecosystem. It will introduce you to Project Tungsten and Catalyst, two of the major advancements of Apache Spark 2.x. You will understand how memory management and binary processing, cache-aware computation, and code generation are used to speed things up dramatically. The book extends to show how to incorporate H2O, SystemML, and DeepLearning4j for machine learning, and Jupyter Notebooks and Kubernetes/Docker for cloud-based Spark. During the course of the book, you will learn about the latest enhancements to Apache Spark 2.x, such as interactive querying of live data and unifying DataFrames and Datasets. You will also learn about the updates on the APIs and how DataFrames and Datasets affect SQL, machine learning, graph processing, and streaming. You will learn to use Spark as a big data operating system, understand how to implement advanced analytics on the new APIs, and explore how easy it is to use Spark in day-to-day tasks. Style and approach This book is an extensive guide to Apache Spark modules and tools and shows how Spark's functionality can be extended for real-time processing and storage with worked examples. **Mastering Cloud Computing Foundations and Applications** Programming Newnes Mastering Cloud Computing is designed for undergraduate students learning to develop cloud computing applications. Tomorrow's applications won't live on a single computer but will be deployed from and reside on a virtual server, accessible anywhere, any time. Tomorrow's application developers need to understand the requirements of building apps for these virtual systems, including concurrent programming, high-performance computing, and data-intensive systems. The book introduces the principles of distributed and parallel computing underlying cloud architectures and specifically focuses on virtualization, thread programming, task programming, and map-reduce programming. There are examples demonstrating all of these and more, with exercises and labs throughout. Explains how to make design choices and tradeoffs to consider when building applications to run in a virtual cloud environment Real-world case studies include scientific, business, and energy-efficiency considerations **Web Services: Concepts, Methodologies, Tools, and Applications** IGI Global Web service technologies are redefining the way that large and small companies are doing business and exchanging information. Due to the critical need for furthering automation, engagement, and efficiency, systems and workflows are becoming increasingly more web-based. Web Services: Concepts, Methodologies, Tools, and Applications is an innovative reference source that examines relevant theoretical frameworks, current practice guidelines, industry standards and standardization, and the latest empirical research findings in web services. Highlighting a range of topics such as cloud computing, quality of service, and semantic web, this multi-volume book is designed for computer engineers, IT specialists, software designers, professionals, researchers, and upper-level students interested in web services architecture, frameworks, and security. **PolyBase Revealed Data Virtualization with SQL Server, Hadoop, Apache Spark, and Beyond** Apress Harness the power of PolyBase data virtualization software to make data from a variety of sources easily accessible through SQL queries while using the T-SQL skills you already know and have mastered. PolyBase Revealed shows you how to use the PolyBase feature of SQL Server 2019 to integrate SQL Server with Azure Blob Storage, Apache Hadoop, other SQL Server instances, Oracle, Cosmos DB, Apache Spark, and more. You will learn how PolyBase can help you reduce storage and other costs by avoiding the need for ETL processes that duplicate data in order to make it

accessible from one source. PolyBase makes SQL Server into that one source, and T-SQL is your golden ticket. The book also covers PolyBase scale-out clusters, allowing you to distribute PolyBase queries among several SQL Server instances, thus improving performance. With great flexibility comes great complexity, and this book shows you where to look when queries fail, complete with coverage of internals, troubleshooting techniques, and where to find more information on obscure cross-platform errors. Data virtualization is a key target for Microsoft with SQL Server 2019. This book will help you keep your skills current, remain relevant, and build new business and career opportunities around Microsoft's product direction. What You Will Learn Install and configure PolyBase as a stand-alone service, or unlock its capabilities with a scale-out cluster Understand how PolyBase interacts with outside data sources while presenting their data as regular SQL Server tables Write queries combining data from SQL Server, Apache Hadoop, Oracle, Cosmos DB, Apache Spark, and more Troubleshoot PolyBase queries using SQL Server Dynamic Management Views Tune PolyBase queries using statistics and execution plans Solve common business problems, including "cold storage" of infrequently accessed data and simplifying ETL jobs Who This Book Is For SQL Server developers working in multi-platform environments who want one easy way of communicating with, and collecting data from, all of these sources **Apache Hadoop YARN Moving beyond MapReduce and Batch Processing with Apache Hadoop 2 Addison-Wesley Professional** "This book is a critically needed resource for the newly released Apache Hadoop 2.0, highlighting YARN as the significant breakthrough that broadens Hadoop beyond the MapReduce paradigm." —From the Foreword by Raymie Stata, CEO of Altiscale The Insider's Guide to Building Distributed, Big Data Applications with Apache Hadoop™ YARN 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. YARN project founder Arun Murthy and project lead Vinod Kumar Vavilapalli demonstrate how YARN increases scalability and cluster utilization, enables new programming models and services, and opens new options beyond Java and batch processing. They walk you through the entire YARN project lifecycle, from installation through deployment. You'll find many examples drawn from the authors' cutting-edge experience—first as Hadoop's earliest developers and implementers at Yahoo! and now as Hortonworks developers moving the platform forward and helping customers succeed with it. Coverage includes YARN's goals, design, architecture, and components—how it expands the Apache Hadoop ecosystem Exploring YARN on a single node Administering YARN clusters and Capacity Scheduler Running existing MapReduce applications Developing a large-scale clustered YARN application Discovering new open source frameworks that run under YARN **Hadoop Beginner's Guide Packt Publishing Ltd** 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. **Mastering Apache Solr 7.x An expert guide to advancing, optimizing, and scaling your enterprise search Packt Publishing Ltd** Accelerate your enterprise search engine and bring relevancy in your search analytics Key Features A practical guide in building expertise with Indexing, Faceting, Clustering and Pagination Master the management and administration of Enterprise Search Applications and services seamlessly Handle multiple data inputs such as JSON, xml, pdf, doc, xls,ppt, csv and much more. Book Description Apache Solr is the only standalone enterprise search server with a REST-like application interface. providing highly scalable, distributed search and index replication for many of the world's largest internet sites. To begin with, you would be introduced to how you perform full text search, multiple filter search, perform dynamic clustering and so on helping you to brush up the basics of Apache Solr. You will also explore the new features and advanced options released in Apache Solr 7.x which will get you numerous performance aspects and making data investigation simpler, easier and powerful. You will learn to build complex queries, extensive filters and how are they compiled in your system to bring relevance in your search tools. You will learn to carry out Solr scoring, elements affecting the document score and how you can optimize or tune the score for the application at hand. You will learn to extract features of documents, writing complex queries in re-ranking the documents. You will also learn advanced options helping you to know what content is indexed and how the extracted content is indexed. Throughout the book, you would go through complex problems with solutions along with varied approaches to tackle your business needs. By the end of this book, you will gain advanced proficiency to build out-of-box smart search solutions for your enterprise demands. What you will learn Design schema using schema API to access data in the database Advance querying and fine-tuning techniques for better performance Get to grips with indexing using Client API Set up a fault tolerant and highly available server with newer distributed capabilities, SolrCloud Explore Apache Tika to upload data with Solr Cell Understand different data operations that can be done while indexing Master advanced querying through Velocity Search UI, faceting and Query Re-ranking, pagination and spatial search Learn to use JavaScript, Python, SolrJ and Ruby for interacting with Solr Who this book is for The book would rightly appeal to developers, software engineers, data engineers and database architects who are building or seeking to build enterprise-wide effective search engines for business intelligence. Prior experience of Apache Solr or Java programming is must to take the best of this book. **Hadoop in Action Simon and Schuster** 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. **Handbook of Research on Social Entrepreneurship and Solidarity Economics IGI Global** Education programs in social entrepreneurship helps to create and fill jobs devoted to developing the local economy, which has become a dual transfer strategy by which a virtuous circle occurs between a retrofitted educational system based on social entrepreneurship, and vocational students who are highly entrepreneurial. The Handbook of Research on Social Entrepreneurship and Solidarity Economics focuses on practical experience and theoretical models for popularizing the concept of social entrepreneurship as a critical element of economic growth. Emphasizing the ways in which social entrepreneurship benefits developing regions, small and medium enterprises, and low-income communities, this handbook of research is a pivotal reference source for professionals, academics, and graduate-level students in the fields of economics, business administration, sociology, education, politics, and international relations. **Mastering Java Machine Learning Packt Publishing Ltd** Become an advanced practitioner with this progressive set of master classes on application-oriented machine learning About This Book Comprehensive coverage of key topics in machine learning with an emphasis on both the theoretical and practical aspects More than 15 open source Java tools in a wide range of techniques, with code and practical usage. More than 10 real-world case studies in machine learning highlighting techniques ranging from data ingestion up to analyzing the results of experiments, all preparing the user for the practical, real-world use of tools and data analysis. Who This Book Is For This book will appeal to anyone with a serious interest in topics in Data Science or those already working in related areas: ideally, intermediate-level data analysts and data scientists with experience in Java. Preferably, you will have experience with the fundamentals of machine learning and now have a desire to explore the area further, are up to grappling with the mathematical complexities of its algorithms, and you wish to learn the complete ins and outs of practical machine learning. What You Will Learn Master key Java machine learning libraries, and what kind of problem each can solve, with theory and practical guidance. Explore powerful techniques in each major category of machine learning such as classification, clustering, anomaly detection, graph modeling, and text mining. Apply machine learning to real-world data with methodologies, processes, applications, and analysis. Techniques and experiments developed around the latest specializations in machine learning, such as deep learning, stream data mining, and active and semi-supervised learning. Build high-performing, real-time, adaptive predictive models for batch- and stream-based big data learning using the latest tools and methodologies. Get a deeper understanding of technologies leading towards a more powerful AI applicable in various domains such as Security, Financial Crime, Internet of Things, social networking, and so on. In Detail Java is one of the main languages used by practicing data scientists; much of the Hadoop ecosystem is Java-based, and it is certainly the language that most production systems in Data Science are written in. If you know Java, Mastering Machine Learning with Java is your next step on the path to becoming an advanced practitioner in Data Science. This book aims to introduce you to an array of advanced techniques in machine learning, including classification, clustering, anomaly detection, stream learning, active learning, semi-supervised learning, probabilistic graph modeling, text mining, deep learning, and big data batch and stream machine learning. Accompanying each chapter are illustrative examples and real-world case studies that show how to apply the newly learned techniques using sound methodologies and the best Java-based tools available today. On completing this book, you will have an understanding of the tools and techniques for building powerful machine learning models to solve data science problems in just about any domain. Style and approach A practical guide to help you explore machine learning—and an array of Java-based tools and frameworks—with the help of practical examples and real-world use cases. **Pro Hadoop Data Analytics Designing and Building Big Data Systems using the Hadoop Ecosystem Apress** Learn advanced analytical techniques and leverage existing tool kits to make your analytic applications more powerful, precise, and efficient. This book provides the right combination of architecture, design, and implementation information to create analytical systems that go beyond the basics of classification, clustering, and recommendation. Pro Hadoop Data Analytics emphasizes best practices to ensure coherent, efficient development. A complete example system will be developed using standard third-party components that consist of the tool kits, libraries, visualization and reporting code, as well as support glue to provide a working and extensible end-to-end system. The book also highlights the importance of end-to-end, flexible, configurable, high-performance data pipeline systems with analytical components as well as appropriate visualization results. You'll discover the importance of mix-and-match or hybrid systems, using different analytical components in one application. This hybrid approach will be prominent in the examples. What You'll Learn Build big data analytic systems with the Hadoop ecosystem Use libraries, tool kits, and algorithms to make development easier and more effective Apply metrics to measure performance and efficiency of components and systems Connect to standard relational databases, noSQL data sources, and more Follow case studies with example components to create your own systems Who This Book Is For Software engineers, architects, and data scientists with an interest in the design and implementation of big data analytical systems using Hadoop, the Hadoop ecosystem, and other associated technologies. **Mastering MongoDB 4.x Expert techniques to run high-volume and fault-tolerant database solutions using MongoDB 4.x, 2nd Edition Packt Publishing Ltd** Leverage the power of MongoDB 4.x to build and administer fault-tolerant database applications Key Features Master the new features and capabilities of MongoDB 4.x Implement advanced data modeling, querying, and administration techniques in MongoDB Includes rich case-studies and best practices followed by expert MongoDB developers Book Description MongoDB is the best platform for working with non-relational data and is considered to be the smartest tool for organizing data in line with business needs. The recently released MongoDB 4.x supports ACID transactions and makes the technology an asset for enterprises across the IT and fintech sectors. This book provides expertise in advanced and niche areas of managing databases (such as modeling and querying databases) along with various administration techniques in MongoDB, thereby helping you become a successful MongoDB expert. The book helps you understand how the newly added capabilities function with the help of some interesting examples and large datasets. You will dive deeper into niche areas such as high-performance configurations, optimizing SQL statements, configuring large-scale sharded clusters, and many more. You will also master best practices in overcoming database failover, and master recovery and backup procedures for database security. By the end of the book, you will have gained a practical understanding of administering database applications both on premises and on the cloud; you will also be able to scale database applications across all servers. What you will learn Perform advanced querying techniques such as indexing and expressions Configure, monitor, and maintain a highly scalable

MongoDB environment Master replication and data sharding to optimize read/write performance Administer MongoDB-based applications on premises or on the cloud Integrate MongoDB with big data sources to process huge amounts of data Deploy MongoDB on Kubernetes containers Use MongoDB in IoT, mobile, and serverless environments Who this book is for This book is ideal for MongoDB developers and database administrators who wish to become successful MongoDB experts and build scalable and fault-tolerant applications using MongoDB. It will also be useful for database professionals who wish to become certified MongoDB professionals. Some understanding of MongoDB and basic database concepts is required to get the most out of this book. **Mastering Apache Cassandra - Second Edition Packt Publishing Ltd** 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. **Machine Learning Algorithms Packt Publishing Ltd** Build strong foundation for entering the world of Machine Learning and data science with the help of this comprehensive guide About This Book Get started in the field of Machine Learning with the help of this solid, concept-rich, yet highly practical guide. Your one-stop solution for everything that matters in mastering the whats and whys of Machine Learning algorithms and their implementation. Get a solid foundation for your entry into Machine Learning by strengthening your roots (algorithms) with this comprehensive guide. Who This Book Is For This book is for IT professionals who want to enter the field of data science and are very new to Machine Learning. Familiarity with languages such as R and Python will be invaluable here. What You Will Learn Acquaint yourself with important elements of Machine Learning Understand the feature selection and feature engineering process Assess performance and error trade-offs for Linear Regression Build a data model and understand how it works by using different types of algorithm Learn to tune the parameters of Support Vector machines Implement clusters to a dataset Explore the concept of Natural Processing Language and Recommendation Systems Create a ML architecture from scratch. In Detail As the amount of data continues to grow at an almost incomprehensible rate, being able to understand and process data is becoming a key differentiator for competitive organizations. Machine learning applications are everywhere, from self-driving cars, spam detection, document search, and trading strategies, to speech recognition. This makes machine learning well-suited to the present-day era of Big Data and Data Science. The main challenge is how to transform data into actionable knowledge. In this book you will learn all the important Machine Learning algorithms that are commonly used in the field of data science. These algorithms can be used for supervised as well as unsupervised learning, reinforcement learning, and semi-supervised learning. A few famous algorithms that are covered in this book are Linear regression, Logistic Regression, SVM, Naive Bayes, K-Means, Random Forest, TensorFlow, and Feature engineering. In this book you will also learn how these algorithms work and their practical implementation to resolve your problems. This book will also introduce you to the Natural Processing Language and Recommendation systems, which help you run multiple algorithms simultaneously. On completion of the book you will have mastered selecting Machine Learning algorithms for clustering, classification, or regression based on for your problem. Style and approach An easy-to-follow, step-by-step guide that will help you get to grips with real-world applications of Algorithms for Machine Learning. **Big Data Computing CRC Press** Due to market forces and technological evolution, Big Data computing is developing at an increasing rate. A wide variety of novel approaches and tools have emerged to tackle the challenges of Big Data, creating both more opportunities and more challenges for students and professionals in the field of data computation and analysis. Presenting a mix of industry cases and theory, Big Data Computing discusses the technical and practical issues related to Big Data in intelligent information management. Emphasizing the adoption and diffusion of Big Data tools and technologies in industry, the book introduces a broad range of Big Data concepts, tools, and techniques. It covers a wide range of research, and provides comparisons between state-of-the-art approaches. Comprised of five sections, the book focuses on: What Big Data is and why it is important Semantic technologies Tools and methods Business and economic perspectives Big Data applications across industries **Optimizing Hadoop for MapReduce Packt Publishing Ltd** This book is an example-based tutorial that deals with Optimizing Hadoop for MapReduce job performance. If you are a Hadoop administrator, developer, MapReduce user, or beginner, this book is the best choice available if you wish to optimize your clusters and applications. Having prior knowledge of creating MapReduce applications is not necessary, but will help you better understand the concepts and snippets of MapReduce class template code.