

## Apache Hive Essentials

*This book takes you on a fantastic journey to discover the attributes of big data using Apache Hive. About This Book 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 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. 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 the 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 In Detail 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 efficiently to find solutions to big data problems Style and approach This book takes on a practical approach which will get you familiarized with Apache Hive and how to use it to efficiently to find solutions to your big data problems. This book covers crucial topics like performance, and data security in order to help you make the most of the Hive working environment. Downloading the example code for this book You can download the example code files for all Packt books you have purchased from your account at <http://www.PacktPub.com>. If you purchased this book elsewhere, you can visit <http://www.PacktPub.com/support> and register to have the files e-ma . . . Set in the post-Civil War Arizona "An Apache Princess" tells the tale of the frontier and the life of Native Americans. It follows the story of two young girls, one being the daughter of the Captain in Arizona fort, the other being the Apache girl, as they go through numerous adventures in the Wild West.*

*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.*

*Solr in Action is a comprehensive guide to implementing scalable search using Apache Solr. This clearly written book walks you through well-documented examples ranging from basic keyword searching to scaling a system for billions of documents and queries. It will give you a deep understanding of how to implement core Solr capabilities. About the Book Whether you're handling big (or small) data, managing documents, or building a website, it is important to be able to quickly search through your content and discover meaning in it. Apache Solr is your tool: a ready-to-deploy, Lucene-based, open source, full-text search engine. Solr can scale across many servers to enable real-time queries and data analytics across billions of documents. Solr in Action teaches you to implement scalable search using Apache Solr. This easy-to-read guide balances conceptual discussions with practical examples to show you how to implement all of Solr's core capabilities. You'll master topics like text analysis, faceted search, hit highlighting, result grouping, query suggestions, multilingual search, advanced geospatial and data operations, and relevancy tuning. This book assumes basic knowledge of Java and standard database technology. No prior knowledge of Solr or Lucene is required. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. What's Inside How to scale Solr for big data Rich real-world examples Solr as a NoSQL data store Advanced multilingual, data, and relevancy tricks Coverage of versions through Solr 4.7 About the Authors Trey Grainger is a director of engineering at CareerBuilder. Timothy Potter is a senior member of the engineering team at LucidWorks. The authors work on the scalability and reliability of Solr, as well as on recommendation engine and big data analytics technologies. Table of Contents PART 1 MEET SOLR Introduction to Solr Getting to know Solr Key Solr concepts Configuring Solr Indexing Text analysis PART 2 CORE SOLR CAPABILITIES Performing queries and handling results Faceted search Hit highlighting Query suggestions Result grouping/field collapsing Taking Solr to production PART 3 TAKING SOLR TO THE NEXT LEVEL SolrCloud Multilingual search Complex query operations Mastering relevancy Get ready to unlock the power of your data. With the fourth edition of 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. Using Hadoop 2 exclusively, author Tom White presents new chapters on YARN and several Hadoop-related projects such as Parquet, Flume, Crunch, and Spark. You'll learn about recent changes to Hadoop, and explore new case studies on Hadoop's role in big data and genomics data processing. Learn fundamental components such as MapReduce, HDFS, and YARN Explore MapReduce in depth, including steps for developing applications with it Set up and maintain a Hadoop cluster running HDFS and MapReduce on YARN Learn two data formats: Avro for data serialization and Parquet for nested data Use data ingestion tools such as Flume (for streaming data) and Sqoop (for bulk data transfer) Understand how high-level data processing tools like Pig, Hive, Crunch, and Spark work with Hadoop Learn the HBase distributed database and the ZooKeeper distributed configuration service Geronimo's Story of His Life Or, Life in Dakota with General Custer Expert techniques for architecting end-to-end big data solutions to get valuable insights Hadoop in Practice Hadoop Operations*

*The Definitive Guide*

*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 efficiently 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.*

*Big Data Science and Analytics is about harnessing the power of data for new insights. The book covers the breadth of activities and methods and tools that Data Scientists use. The content focuses on concepts, principles and practical applications that are applicable to any industry and technology environment, and the learning is supported and explained with examples that you can replicate using open-source software. This book will help you: Become a contributor on a data science team Deploy a structured lifecycle approach to data analytics problems Apply appropriate analytic techniques and tools to analyzing big data Learn how to tell a compelling story with data to drive business action Prepare for EMC Proven Professional Data Science Certification Corresponding data sets are available from the book's page at Wiley which you can find on the Wiley site by searching for the ISBN 978111876138. Get started discovering, analyzing, visualizing, and presenting data in a meaningful way today!*

*Microsoft Azure Essentials from Microsoft Press is a series of free ebooks designed to help you advance your technical skills with Microsoft Azure. The first ebook in the series, Microsoft Azure Essentials: Fundamentals of Azure, introduces developers and IT professionals to the wide range of capabilities in Azure. The authors – both Microsoft MVPs in Azure – present both conceptual and how-to content for key areas, including: Azure Websites and Azure Cloud Services Azure Virtual Machines Azure Storage Azure Virtual Networks Databases Azure Active Directory Management tools Business scenarios Watch Microsoft Press's blog and Twitter (@MicrosoftPress) to learn about other free ebooks in the "Microsoft Azure Essentials" series. Instant Apache Hive Essentials How-to*

*Hadoop: The Definitive Guide*

*A Complete Introduction*

*Practical Enterprise Data Lake Insights*

*A Guide for Developers and Administrators*

*Apache Oozie Essentials*

*Understanding Gang of Four, architectural, functional, and reactive design patterns and how to implement them on modern Java platforms, such as java 12 and beyond Key FeaturesLearn OOP, functional, and reactive patterns for creating readable and maintainable codeExplore architectural patterns and practices for building scalable and reliable applicationsTackle all kinds of performance-related issues and streamline development using design patternsBook Description Java design patterns are reusable and proven solutions to software design problems. This book covers over 60 battle-tested design patterns used by developers to create functional, reusable, and flexible software. Hands-On Design Patterns with Java starts with an introduction to the Unified Modeling Language (UML), and delves into class and object diagrams with the help of detailed examples. You'll study concepts and approaches to object-oriented programming (OOP) and OOP design patterns to build robust applications. As you advance, you'll explore the categories of GOF design patterns, such as behavioral, creational, and structural, that help you improve code readability and enable large-scale reuse of software. You'll also discover how to work effectively with microservices and serverless architectures by using cloud design patterns, each of which is thoroughly explained and accompanied by real-world programming solutions. By the end of the book, you'll be able to speed up your software development process using the right design patterns, and you'll be comfortable working on scalable and maintainable projects of any size. What you will learnUnderstand the significance of design patterns for software engineeringVisualize software design with UML diagramsStrengthen your understanding of OOP to create reusable software systemsDiscover GOF design patterns to develop scalable applicationsExamine programming challenges and the design patterns that solve themExplore architectural patterns for microservices and cloud developmentWho this book is for If you are a developer who*

wants to learn how to write clear, concise, and effective code for building production-ready applications, this book is for you. Familiarity with the fundamentals of Java is assumed.

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

*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, scalable 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.*

*Get Started Fast with Apache Hadoop® 2, YARN, and Today's Hadoop Ecosystem With Hadoop 2.x and YARN. Hadoop moves beyond MapReduce to become practical for virtually any type of data processing. Hadoop 2.x and the Data Lake concept represent a radical shift away from conventional approaches to data usage and storage. Hadoop 2.x installations offer unmatched scalability and breakthrough extensibility that supports new and existing Big Data analytics processing methods and models. Hadoop® 2 Quick-Start Guide is the first easy, accessible guide to Apache Hadoop 2.x, YARN, and the modern Hadoop ecosystem. Building on his unsurpassed experience teaching Hadoop and Big Data, author Douglas Edline covers all the basics you need to know to install and use Hadoop 2 on personal computers or servers, and to navigate the powerful technologies that complement it. Edline concisely introduces and explains every key Hadoop 2 concept, tool, and service, illustrating each with a simple "beginning-to-end" example and identifying trustworthy, up-to-date resources for learning more. This guide is ideal if you want to learn about Hadoop 2 without getting mired in technical details. Douglas Edline will bring you up to speed quickly, whether you're a user, admin, devops specialist, programmer, architect, analyst, or data scientist. Coverage includes Understanding Hadoop 2 architecture and how they improve on Hadoop 1 with MapReduce Understanding Hadoop-based Data Lakes versus RDBMS Data Warehouses Installing Hadoop 2 and core services on Linux machines, virtualized sandboxes, or clusters Exploring the Hadoop Distributed File System (HDFS) Understanding the essentials of MapReduce and YARN application programming Simplifying programming and data movement with Apache Pig, Hive, Sqoop, Flume, Oozie, and HBase Observing application progress, controlling jobs, and managing workflows Managing Hadoop efficiently with Apache Ambari—including recipes for HDFS-to NFSv3 gateway, HDFS snapshots, and YARN configuration Learning basic Hadoop 2 troubleshooting, and installing Apache Hue and Apache Spark Use this practical guide to successfully handle the challenges encountered when designing an enterprise data lake and learn industry best practices to resolve issues. When designing an enterprise data lake you often hit a roadblock when you must leave the comfort of the relational world and learn the nuances of handling non-relational data. Starting from sourcing data into the Hadoop ecosystem, you will go through stages that can bring up tough questions such as data processing, data querying, and security. Concepts such as change data capture and data streaming are covered. The book takes an end-to-end solution approach in a data lake environment that includes data security, high availability, data processing, data streaming, and more. Each chapter includes application of a concept, code snippets, and use case demonstrations to provide you with a practical approach. You will learn the concept, scope, application, and starting point. What You'll Learn Get to know data lake architecture and design principles Implement data capture and streaming strategies Implement data processing strategies in Hadoop Understand the data lake security framework and availability model Who This Book Is For Big data architects and solution architects*

*Apache Hadoop YARN*

*Apache Hive Cookbook*

*Programming Hive*

*Trino: The Definitive Guide*

*A Guide to Hadoop's Data Warehouse System*

*Big Data in Practice*

*If your organization is looking for a storage solution to accommodate a virtually endless amount of data, this book will show you how Apache HBase can fulfill your needs. As the open source implementation of Google's BigTable architecture, HBase scales to billions of rows and millions of columns, while ensuring that write and read performance remain constant.HBase: The Definitive Guideprovides the details you require, whether you simply want to evaluate this high-performance, non-relational database, or put it into practice right away. HBase's adoption rate is beginning to climb, and several IT executives are asking pointed questions about this high-capacity database. This is the only book available to give you meaningful answers. Learn how to distribute large datasets across an inexpensive cluster of commodity servers Develop HBase clients in many programming languages, including Java, Python, and Ruby Get details on HBase's primary storage system, HDFS—Hadoop ' s distributed and replicated filesystem Learn how HBase's native interface to Hadoop ' s MapReduce framework enables easy development and execution of batch jobs that can scan entire tables Discover the integration between HBase and other facets of the Apache Hadoop project*

*A comprehensive guide to design, build and execute effective Big Data strategies using Hadoop Key Features -Get an in-depth view of the Apache Hadoop ecosystem and an overview of the architectural patterns pertaining to the popular Big Data platform -Conquer different data processing and analytics challenges using a multitude of tools such as Apache Spark, Elasticsearch, Tableau and more -A comprehensive, step-by-step guide that will teach you everything you need to know, to be an expert Hadoop Architect Book Description The complex structure of data these days requires sophisticated solutions for data transformation, to make the information more accessible to the users.This book empowers you to build such solutions with relative ease with the help of Apache Hadoop, along with a host of other Big Data tools. This book will give you a complete understanding of the data lifecycle management with Hadoop, followed by modeling of structured and unstructured data in Hadoop. It will also show you how to design real-time streaming pipelines by leveraging tools such as Apache Spark, and build efficient enterprise search solutions using Elasticsearch. You will learn to build enterprise-grade analytics solutions on Hadoop, and how to visualize your data using tools such as Apache Superset. This book also covers techniques for deploying your Big Data solutions on the cloud Apache Ambari, as well as expert techniques for managing and administering your Hadoop cluster. By the end of this book, you will have all the knowledge you need to build expert Big Data systems. What you will learn Build an efficient enterprise Big Data strategy centered around Apache Hadoop Gain a thorough understanding of using Hadoop with various Big Data frameworks such as Apache Spark, Elasticsearch and more Set up and deploy your Big Data environment on premises or on the cloud with Apache Ambari Design effective streaming data pipelines and build your own enterprise search solutions Utilize historical data to build your analytics solutions and visualize them using popular tools such as Apache Superset Plan, set up and administer your Hadoop cluster efficiently Who this book is for This book is for Big Data professionals who want to fast-track their career in the Hadoop industry and become an expert Big Data architect. Project managers and mainframe professionals looking forward to build a career in Big Data Hadoop will also find this book to be useful. Some understanding of Hadoop is required to get the best out of this book.*

*BRIDGES THE GAP BETWEEN SAS AND R, ALLOWING USERS TRAINED IN ONE LANGUAGE TO EASILY LEARN THE OTHER SAS and R are widely-used, very different software environments. Prized for its statistical and graphical tools, R is an open-source programming language that is popular with statisticians and data miners who develop statistical software and analyze data. SAS (Statistical Analysis System) is the leading corporate software in analytics thanks to its faster data handling and smaller learning curve. SAS For R Users enables entry-level data scientists to take advantage of the best aspects of both tools by providing a cross-functional framework for users who already know R but may need to work with SAS. Those with knowledge of both R and SAS are of far greater value to employers, particularly in corporate settings. Using a clear, step-by-step approach, this book presents an analytics workflow that mirrors that of the everyday data scientist. This up-to-date guide is compatible with the latest R packages as well as SAS University Edition. Useful for anyone seeking employment in data science, this book: Instructs both practitioners and students fluent in one language seeking to learn the other Provides command-by-command translations of R to SAS and SAS to R Offers examples and applications in both R and SAS Presents step-by-step guidance on workflows, color illustrations, sample code, chapter quizzes, and more Includes sections on advanced methods and applications Designed for professionals, researchers, and students. SAS For R Users is a valuable resource for those with some knowledge of coding and basic statistics who wish to enter the realm of data science and business analytics. AJAY OHRI is the founder of analytics startup DecisionStats.com. His research interests include spreading open source analytics, analyzing social media manipulation with mechanism design, simpler interfaces to cloud computing, investigating climate change, and knowledge flows. He currently advises startups in analytics off shoring, analytics services, and analytics. He is the author of Python for R Users: A Data Science Approach (Wiley), R for Business Analytics, and R for Cloud Computing.*

*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 Matej 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 and regression*

*Language big data add value to your business Social media APIs, web-tracking, and other technologies help companies acquire and handle massive amounts of data to better understand their customers, products, competition, and markets. Armed with the insights from big data, companies can improve customer experience and products, add value, and increase return on investment. The tricky part for busy IT professionals and executives is how to get this done, and that's where this practical book comes in. Big Data: Understanding How Data Powers Big Business is a complete how-to guide to leveraging big data to drive business value. Full of practical techniques, real-world examples, and hands-on exercises, this book explores the technologies involved, as well as how to find areas of the organization that can take full advantage of big data. Shows how to decompose current business practices in order to link big data initiatives to the organization ' s value creation processes Explores different value creation processes and models Explains issues surrounding operationalizing big data, including organizational structures, education challenges, and new big data-related roles Provides methodology worksheets and exercises to readers can apply techniques Includes real-world examples from a variety of organizations leveraging big data Big Data: Understanding How Data Powers Big Business is written by one of Big Data's preminent experts, William Schmarzo. Don't miss his invaluable insights and advice.*

*Microsoft Azure Essentials - Fundamentals of Azure*

*A Book for Data Scientists*

*Moving Beyond MapReduce and Batch Processing with Apache Hadoop 2*

*Big Data Processing Made Simple*

*Big Data Processing Made Easy*

*Big Data*

*Apache Hive EssentialsPackt Publishing Ltd*

*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 Kjerungaard, 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. 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. Easy, hands-on recipes to help you understand Hive and its integration with frameworks that are used widely in today's big data world About This Book Grasp a complete reference of different Hive topics. Get to know the latest recipes in development in Hive including CRUD operations Understand Hive internals and integration of Hive with different frameworks used in today's world. Who This Book Is For The book is intended for those who want to start in Hive or who have basic understanding of Hive framework. Prior knowledge of basic SQL command is also required What You Will Learn Learn different features and offering on the latest Hive Understand the working and structure of the Hive internals Get an insight on the latest development in Hive Framework Grasp the concepts of Hive Data Model Master the key concepts like Partition, Buckets and Statistics Know how to integrate Hive with other frameworks such as Spark, Accumulo, etc In Detail Hive was developed by Facebook and later open sourced in Apache community. Hive provides SQL like interface to run queries on Big Data frameworks. Hive provides SQL like syntax also called as HiveQL that includes all SQL capabilities like analytical functions which are the need of the hour in today's Big Data world. This book provides you easy installation steps with different types of metastores supported by Hive. This book has simple and easy to learn recipes for configuring Hive clients and services. You would also learn different Hive optimizations including Partitions and Bucketing. The book also covers the source code explanation of latest Hive version. Hive Query Language is being used by other frameworks including spark. Towards the end you will cover integration of Hive with these frameworks. Style and approach Starting with the basics and covering the core concepts with the practical usage, this book is a complete guide to learn and explore Hive offerings.*

*Unleash the power of Apache Oozie to create and manage your big data and machine learning pipelines in one go About This Book Teaches you everything you need to know to get started with Apache Oozie from scratch and manage your data pipelines effortlessly Learn to write data ingestion workflows with the help of real-life examples from the author's own personal experience Embed Spark jobs to run your machine learning models on top of Hadoop Who This Book Is For If you are an expert Hadoop user who wants to use Apache Oozie to handle workflows efficiently, this book is for you. This book will be handy to anyone who is familiar with the basics of Hadoop and wants to automate data and machine learning pipelines. What You Will Learn Install and configure Oozie from source code on your Hadoop cluster Dive into the world of Oozie with Java MapReduce jobs Schedule Hive ETL and data ingestion jobs Import data from a database through Sqoop jobs in HDFS Create and process data pipelines with Pig, Hive scripts as per business requirements. Run machine learning Spark jobs on Hadoop Create quick Oozie jobs using Hue Make the most of Oozie's security capabilities by configuring Oozie's security in Detail As more and more organizations are discovering the use of big data analytics, interest in platforms that provide storage, computation, and analytic capabilities is booming exponentially. This calls for data management. Hadoop caters to this need. Oozie fulfils this necessity for a scheduler for a Hadoop job by acting as a cron to better analyze data. Apache Oozie Essentials starts off with the basics right from installing and configuring Oozie from source code on your Hadoop cluster to managing your complex clusters. You will learn how to create data ingestion and machine learning workflows. This book is sprinkled with the examples and exercises to help you take your big data learning to the next level. You will discover how to write workflows to run your MapReduce, Pig ,Hive, and Sqoop scripts and schedule them to run at a specific time or for a specific business requirement using a coordinator. This book has engaging real-life exercises and examples to get you in the thick of things. Lastly, you'll get a grip of how to embed Spark jobs, which can be used to run your machine learning models on Hadoop. By the end of the book, you will have a good knowledge of Apache Oozie. You will be capable of using Oozie to handle large Hadoop workflows and even improve the availability of your Hadoop environment. Style and approach This book is a hands-on guide that explains Oozie using real-world examples. Each chapter is blended beautifully with fundamental concepts sprinkled in-between case study solution algorithms and topped off with self-learning exercises.*

*Storage and Analysis at Internet Scale*

*Solr in Action*

*Hadoop in Action*

*Practical Data Science with Hadoop and Spark*

*An Apache Princess*

*Essential techniques to help you process, and get unique insights from, big data, 2nd Edition*

*"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™ Hadoop in Action technical leaders show you how to develop new applications and adapt existing code to fully leverage these revolutionary advances." -- From the Amazon*

*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 with a detailed look at the core components of 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. 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*

*Inspired by practical, step-by-step instructions and clear explanations for the most important and useful tasks. This book provides quick recipes for using Hive to read data in various formats, efficiently querying this data, and extending Hive with any custom functions you may need to insert your own logic into the data pipeline. This book is written for data analysts and developers who want to use their current knowledge of SQL to be more productive with Hadoop. It assumes that readers are comfortable writing SQL queries and are familiar with Hadoop at the level of the classic WordCount example. 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 large-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*

*Practical Hadoop Ecosystem*

*The Linux Command Line, 2nd Edition*

*Apache Hive Essentials*

*Learn design patterns that enable the building of large-scale software architectures*

*You've experienced the shiny, point-and-click surface of your Linux computer—now dive below and explore its depths with the power of the command line. The Linux Command Line takes you from your very first terminal keystrokes to writing full programs in Bash, the most popular Linux shell (or command line). Along the way you'll learn the timeless skills handed down by generations of experienced, mouse-shunning gurus: file naming conventions, pattern matching with regular expressions, and more. In addition to that practical knowledge, author William Shotts reveals the philosophy behind these tools and the rich heritage that your desktop Linux machine has inherited from Unix supercomputers of yore. As you make your way through the book's short, easily-digestible chapters, you'll learn how to: • Create and delete files, directories, and symlinks • Administer your system, including networker, package installation, and process management • Use standard input and output, redirection, and pipelines • Edit files with Vi, the world's most popular text editor • Write shell scripts to automate common or boring tasks • Slice and dice text files with cut, paste, grep, patch, and sed Once you overcome your initial "shell shock," you'll find that the command line is a natural and expressive way to communicate with your computer. Just don't be surprised if your mouse starts to gather dust. Oklahoma S.M Barret wrote down and edited Apache Chief Geronimo's story of his life. Learn to use Apache Pig to develop lightweight big data applications easily and quickly. This book shows you many optimization techniques and covers every context where Pig is used in big data analytics. Beginning Apache Pig shows you how Pig is easy to learn and requires relatively little time to develop big data applications.The book is divided into four parts: the complete features of Apache Pig; integration with other tools; how to solve complex business problems; and optimization of tools.You'll discover topics such as MapReduce which you try it cannot meet every business need; the features of Pig Latin such as data types for each load, store, joins, groups, and ordering; how Pig workflows can be created; submitting Pig jobs using Hue; and working with Oozie. You'll also see how to extend the framework by writing UDFs and custom load, store, and filter functions. Finally you'll cover different optimization techniques such as gathering statistics about a Pig script, joining strategies, parallelism, and the role of data formats in good performance. What You Will Learn- Use all the features of Apache Pig- Integrate Apache Pig with other tools- Extend Apache Pig- Optimize Pig Latin code- Solve different use cases for Pig LatinWho This Book Is ForAll levels of IT professionals.*

*The Complete Guide to Data Science with Hadoop—For Technical Professionals, Businesspeople, and Students Demand is soaring for professionals who can solve real data science problems with Hadoop and Spark. Practical Data Science with Hadoop® and Spark is your complete guide to doing just that. Drawing on immense experience with Hadoop and big data, three leading experts bring together everything you need: high-level concepts, deep-dive techniques, real-world use cases, practical applications, and hands-on tutorials. The authors introduce the essentials of data science and the modern Hadoop ecosystem, explaining how Hadoop and Spark have evolved into an effective platform for solving data science problems at scale. In addition to comprehensive application coverage, the authors also provide useful guidance on the important steps of data ingestion, data munging, and visualization. Once the groundwork is in place, the authors focus on specific applications, including machine learning, predictive modeling for sentiment analysis, clustering for document analysis, anomaly detection, and natural language processing (NLP). This guide provides a strong technical foundation for those who want to do practical data science, and also presents business-driven guidance on how to apply Hadoop and Spark to optimize ROI of data science initiatives. Learn what data science is, how it has evolved, and how to plan a data science career How data volume, variety, and velocity shape data science use cases Hadoop and its ecosystem, including HDFS, MapReduce, YARN, and Spark Data importation with Hive and Spark Data quality, preprocessing, preparation, and modeling Visualization: surfacing insights from huge data sets Machine learning: classification, regression, clustering, and anomaly detection Algorithms and Hadoop tools for predictive modeling Cluster analysis and similarity functions Large-scale anomaly detection NLP: applying data science to human language*

*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*

*Hands-On Design Patterns with Java*

*Over 60 recipes for implementing big data processing and analytics using Apache Spark and Python*

*Hive 2 Quick-Start Guide*

*Modern Big Data Processing with Hadoop*



Hadoop Essentials

Spark: The Definitive Guide

The best-selling author of Big Data is back, this time with a unique and in-depth insight into how specific companies use big data. Big data is on the tip of everyone's tongue. Everyone understands its power and importance, but many fail to grasp the actionable steps and resources required to utilise it effectively. This book fills the knowledge gap by showing how major companies are using big data every day, from an up-close, on-the-ground perspective. From technology, media and retail, to sport teams, government agencies and financial institutions, learn the actual strategies and processes being used to learn about customers, improve manufacturing, spur innovation, improve safety and so much more. Organised for easy dip-in navigation, each chapter follows the same structure to give you the information you need quickly. For each company profiled, learn what data was used, what problem it solved and the processes put it place to make it practical, as well as the technical details, challenges and lessons learned from each unique scenario. Learn how predictive analytics helps Amazon, Target, John Deere and Apple understand their customers Discover how big data is behind the success of Walmart, LinkedIn, Microsoft and more Learn how big data is changing medicine, law enforcement, hospitality, fashion, science and banking Develop your own big data strategy by accessing additional reading materials at the end of each chapter

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

Find the right big data solution for your business organization Big data management is one of the major challenges facing business, industry, and not-for-profit organizations. Data sets such as customer transactions for a mega-retailer, weather patterns monitored by meteorologists, or social network activity can quickly outpace the capacity of traditional data management tools. If you need to develop or manage big data solutions, you'll appreciate how these four experts define, explain, and guide you through this new and often confusing concept. You'll learn what it is, why it matters, and how to choose and implement solutions that work. Effectively managing big data is an issue of growing importance to businesses, not-for-profit organizations, government, and IT professionals. Authors are experts in information management, big data, and a variety of solutions. Explains big data in detail and discusses how to select and implement a solution, security concerns to consider, data storage and presentation issues, analytics, and much more. Provides essential information in a no-nonsense, easy-to-understand style that is empowering.

Big Data For Dummies cuts through the confusion and helps you take charge of big data solutions for your organization.

Can we add value to the current Apache Hive decision-making process (largely qualitative) by incorporating uncertainty modeling (more quantitative)? Apache Hive in management - Strategic planning How will the Apache Hive team and the organization measure complete success of Apache Hive? Will Apache Hive deliverables need to be tested and, if so, by whom? Who will be responsible for deciding whether Apache Hive goes ahead or not after the initial investigations? This premium Apache Hive self-assessment will make you the credible Apache Hive domain auditor by revealing just what you need to know to be fluent and ready for any Apache Hive challenge. How do I reduce the effort in the Apache Hive work to be done to get problems solved? How can I ensure that plans of action include every Apache Hive task and that every Apache Hive outcome is in place? How will I save time investigating strategic and tactical options and ensuring Apache Hive costs are low? How can I deliver tailored Apache Hive advice instantly with structured going-forward plans? There's no better guide through these mind-expanding questions than acclaimed best-selling author Gerard Blokdyk. Blokdyk ensures all Apache Hive essentials are covered, from every angle: the Apache Hive self-assessment shows succinctly and clearly that what needs to be clarified to organize the required activities and processes so that Apache Hive outcomes are achieved. Contains extensive criteria grounded in past and current successful projects and activities by experienced Apache Hive practitioners. Their mastery, combined with the easy elegance of the self-assessment, provides its superior value to you in knowing how to ensure the outcome of any efforts in Apache Hive are maximized with professional results. Your purchase includes access details to the Apache Hive self-assessment dashboard download which gives you your dynamically prioritized projects-ready tool and shows you exactly what to do next. Your exclusive instant access details can be found in your book. You will receive the following contents with New and Updated specific criteria: - The latest quick edition of the book in PDF - The latest complete edition of the book in PDF, which criteria correspond to the criteria in... - The Self-Assessment Excel Dashboard, and... - Example pre-filled Self-Assessment Excel Dashboard to get familiar with results generation ...plus an extra, special, resource that helps you with project managing. INCLUDES LIFETIME SELF ASSESSMENT UPDATES Every self assessment comes with Lifetime

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

Understanding How Data Powers Big Business  
Big Data For Dummies  
Discovering, Analyzing, Visualizing and Presenting Data  
Learning Spark  
How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results

HBase  
Combine the power of Apache Spark and Python to build effective big data applications Key Features Perform effective data processing, machine learning, and analytics using PySpark Overcome challenges in developing and deploying Spark solutions using Python Explore recipes for efficiently combining Python and Apache Spark to process data Book Description Apache Spark is an open source framework for efficient cluster computing with a strong interface for data parallelism and fault tolerance. The PySpark Cookbook presents effective and time-saving recipes for leveraging the power of Python and putting it to use in the Spark ecosystem. You'll start by learning the Apache Spark architecture and how to set up a Python environment for Spark. You'll then get familiar with the modules available in PySpark and start using them effortlessly. In addition to this, you'll discover how to abstract data with RDDs and DataFrames, and understand the streaming capabilities of PySpark. You'll then move on to using ML and MLlib in order to solve any problems related to the machine learning capabilities of PySpark and use GraphFrames to solve graph-processing problems. Finally, you will explore how to deploy your applications to the cloud using the spark-submit command. By the end of this book, you will be able to use the Python API for Apache Spark to solve any problems associated with building data-intensive applications. What you will learn Configure a local instance of PySpark in a virtual environment Install and configure Jupyter in local and multi-node environments Create DataFrames from JSON and a dictionary using pyspark.sql Explore regression and clustering models available in the ML module Use DataFrames to transform data used for modeling Connect to PubNub and perform aggregations on streams Who this book is for The PySpark Cookbook is for you if you are a Python developer looking for hands-on recipes for using the Apache Spark 2.x ecosystem in the best possible way. A thorough understanding of Python (and some familiarity with Spark) will help you get the best out of the book.

Boots and Saddles  
Practical Hive  
Designing and Building Effective Analytics at Scale  
SAS for R Users  
A Definitive Guide to Hadoop-Related Frameworks and Tools  
Handle Data-Driven Challenges in an Enterprise Big Data Lake