Hive Sql For Hadoop | 785fb2ad9b21623c28f7bf44167bb35d


This guide is an ideal learning tool and reference for Apache Pig, the programming language that helps programmers describe and run large data projects on Hadoop. With Pig, they can analyze data without having to create a full-fledged application—making it easy for them to experiment with new data sets. 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. 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.

Describes the features and functions of Apache Hive, the data infrastructure for Hadoop. 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 in action. In action gives practical advice for choosing the right framework configuration at design time. 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. If you are a data analyst, developer, or simply someone who wants to use Hive to explore and analyze data in Hadoop, this book is 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 related features of the database Hive. This book 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 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 the data stored in Hadoop. Also 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. 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 2.3.1 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, along with 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 in 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 that 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. This book presents unique techniques to conquer different Big Data processing and analytics challenges using Hadoop. Practical examples are provided to boost your understanding of Big Data concepts and their implementation. By the end of the book, you will have all the knowledge and skills you need to become a true Big Data expert. A practical guide to implementing your enterprise data lake using Lambda Architecture as the base About This Book Build a full-fledged data lake for your organization with popular big data technologies using the Lambda architecture as the base Deliver into the big data technology stack needed to meet your business needs by using a combination of Hadoop and modern data platforms Use Hadoop to enable your business to scale and transform your data lake for enterprise-wide use-cases. Who This Book Is For Java developers and architects who would like to implement a data lake for their enterprise will find this book useful. If you want to get hands-on experience with the Lambda Architecture and big data technologies by implementing a practical solution using these technologies, this book will also help you. What You Will Learn Build an enterprise-level data lake using the relevant big data technologies Understand the core of the Lambda architecture and how to apply it in an enterprise Learn the technical details around Sqoop and its functionalities Integrate Kafka with Hadoop components to acquire enterprise data Use flume with streaming technologies for stream-based processing Understand stream-based processing with reference to Apache Spark Streaming Incorporate Hadoop components and know the advantages they provide for enterprise data lakes Build fast, streaming, and high-performance applications using Elasticsearch Make your data ingestion process consistent across various data formats with configurability Process your data to derive intelligence using machine learning algorithms In Detail The term “Data Lake” has recently emerged as a prominent term in the big data industry. Data scientists can make use of it in deriving meaningful insights that can be used by businesses to redefine or transform the way they operate. Lambda architecture is also emerging as one of the very eminent patterns in the big data landscape, as it not only helps to derive useful information from historical data but also correlates real-time data to enable business to take critical decisions. This book tries to bring these two important aspects – data lake and lambda architecture—together. This book is divided into three main sections. The first introduces you to the concept of data lakes, the importance of data lakes in enterprises, and getting you up-to-speed with the Lambda architecture. The second section delves into the principal components of building a data lake using the Lambda architecture. It introduces you to popular big data technologies such as Apache Hadoop, Spark, Sqoop, Flume, and Elasticsearch. The third section is a highly practical demonstration of putting it all together, and shows you how an enterprise data lake can be implemented with several real-world use-cases. By the end of this book, you will be able to choose the right big data technologies using the lambda architectural patterns to build your enterprise data lake style. Approach The book takes a pragmatic approach, showing ways to leverage big data technologies and lambda architecture to build an enterprise-level data lake. Learn how to write, tune, and port SQL queries and other statements for a Big Data environment, using Impala—the massively parallel processing SQL query engine for Apache Hadoop. The best practices in this practical guide help you design database schemas that not only interoperate with other Hadoop components, and are convenient for administrators to manage and monitor, but also accommodate future expansion in data size and evolution of software capabilities. Written by John Russell, documentation lead for the Cloudera Impala project, this book gets you working with the most recent Impala releases quickly. Ideal for database developers and business analysts, the latest revision covers analytics functions, complex types, incremental statistics, and joins, and submission of jobs to the Impala server. Getting Started with Impala includes advice from the Cloudera development team, as well as insights from its consulting engagements with customers. Learn how Impala integrates with a wide range of Hadoop components Attain high performance and scalability for huge data sets on production clusters Explore common developer tasks, such as porting code to Impala and optimizing performance Use tutorials for working with billion-row tables, date- and time-based values, and other techniques Learn how to transition from rigid schemas to a flexible model that evolves as needs change. Take a deep dive into joins and the roles of statistics Find the right big data solution for your business or organization Big data management is one of the major challenges facing industry, business, 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 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 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. This book introduces you to the Big Data processing techniques addressing but not limited to various BI (business intelligence) requirements, such as reporting, batch analytics, online analytical processing (OLAP), data mining and Warehousing, and predictive analytics. The book has been written on IBMs Platform of Hadoop framework. IBM Infosphere BigInsight has the highest amount of tutorial matter available free of cost on Internet and makes it easy to acquire proficiency in this technique. This therefore
becomes highly vulnerable coaching materials in easy to learn steps. The book optimally provides the courseware as per MCA and M. Tech Level Syllabi of most of the Universities. All components of big data platform like HDP, Hive Pig, Sqoop, Flume , Hadoop Streaming, Oozie: HBase, HDFS, FlumeNG, Whirr, Cloudera, Fuse , Zookeeper and Mahout: Machine learning for Hadoop has been discussed in sufficient Detail with hands on Exercises on each.Get expert guidance on architecting end-to-end data management solutions with Apache Hadoop. While many sources explain how to use various components in the Hadoop ecosystem, this practical book takes you through architectural considerations necessary to tie those components together into a complete tailored application, based on your particular use case. To reinforce those lessons, the book’s second section provides detailed examples of architectures used in some of the most commonly found Hadoop applications. Whether you’re designing a new Hadoop application, or planning to integrate Hadoop into your existing data infrastructure, Hadoop Application Architectures will skillfully guide you through the process. This book covers: Factors to consider when using Hadoop to store and model data Best practices for moving data in and out of the system Data processing frameworks, including MapReduce, Spark, and Hive Common Hadoop processing patterns, such as removing duplicate records and using windowing analytics Giraph, GraphX, and other tools for large graph processing on Hadoop Using workflow orchestration and scheduling tools such as Apache Oozie Near-real-time stream processing with Apache Storm, Apache Spark Streaming, and Apache Flume Apache Hadoop provides for enterprise Hadoop. Second Edition updated, 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 3.0. 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 MANAGEMENT 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 Let Hadoop For Dummies help harness the power of Hadoop and raise the informational power of your organization. Learn how 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, this how-to has something to help you with. Hadoop gives you the world’s best tool for solving problems 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. Apache Hive is the new member in the database family that works within the Hadoop ecosystem. It provides all great features like data summarization, ad-hoc query, and analysis of large datasets. If you are not a good programmer, then this edition will teach you how to use hive queries without writing complex codes. Most users face the problem of not getting a dedicated course on Hive. The goal of this e-book is to cater everything about Hive and only Hive with minimum jargons. The notes, lessons and hands-on examples in this small e-book are simplified and tactfully presented to solve all your Hive queries. Instead of writing long code for MapReduce or Java, the e-book shows tips on writing the same program with a minimum code snippet. Beginners as well as peers will thoroughly enjoy this book. They will discover and learn more hive patterns for data processing and data integrations. Unlike other e-book, where they skip basic detail thinking users having prior subject knowledge. This edition has given complete attention to ensure every small aspect of the hive like “how to set up and configure Hive in your
environment”. This e-book is also helpful for those who just want to explore Hive and don’t want to spend big bucks for short courses. You will quickly learn, apply and share your Hive knowledge with this e-book. Table of Content Chapter 1: Introduction 1. What is Hive? 2. Hive Architecture 3. Different modes of Hive 4. What is Hive Server2 (HS2)? 5. Hive vs Map Reduce Chapter 2: Installation and Configuration 1. Installation of Hive 2. Hive shell commands 3. Install and configure MYSQL database Chapter 3: Data operations 1. Data types in Hive 2. Creation and dropping of Database in Hive 3. Create, Drop and altering of tables in Hive 4. Table types and its Usage 5. Partitions 6. Buckets Chapter 4: Queries and Implementation 1. Order by query 2. Group by query 3. Sort by 4. Cluster By 5. Distribute By 6. Join queries 7. Different type of joins 8. Sub queries 9. Embedding custom scripts 10. UDFs (User Define Functions) Chapter 5: Query Language, Built-in Operators and Functions 1. Hive Query Language (HQL) 2. Built-in operators 3. Built-in functions Chapter 6: Data Extraction 1. Working with Structured Data using Hive 2. Working with Semi structured data using Hive (XML, JSON) 3. Hive in Real time projects - When and Where to UseThis book will be a step-by-step tutorial, which practically teaches working with big data on SQL Server through sample examples in increasing complexity. Microsoft SQL Server 2012 with Hadoop is specifically targeted at readers who want to cross-pollinate their Hadoop skills with SQL Server 2012 business intelligence and data analytics. A basic understanding of traditional RDBMS technologies and query processing techniques is essentially looking for a scalable streaming big data solution. This book provides meaningful answers, whether you’re evaluating this non-relational data platform or wanting to put it into practice right away. Discover how Hive and Hadoop make it easier to distribute large datasets across an inexpensive cluster of commodity servers. Access HBase with native Java clients, or with gateway servers providing REST, Avro, or Thrift APIs. Get details on HBase’s architecture, including the storage format, write-ahead log, background processes, and more. Integrate HBase with Hadoop’s MapReduce framework for massively parallelized data processing jobs. Learn how to tune clusters, design schemas, copy tables, import bulk data, decommission nodes, and many other tasks. This book is a practical guide on using the Apache Hadoop projects including MapReduce, HDFS, Apache Hive, Apache HBase, Apache Kafka, Apache Mahout and Apache Solr. From setting up the environment to running sample applications each chapter is a practical tutorial on using a Apache Hadoop ecosystem project. While several books on Apache Hadoop are available, most are based on the main projects MapReduce and HDFS and none discusses the other Apache Hadoop ecosystem projects and how these all work together as a cohesive big data development platform. What you’ll learn How to set up environment in Linux for Hadoop projects using Cloudera Hadoop Distribution CDH 5. How to run a MapReduce job How to store data with Apache Hive, Apache HBase How to index data in HDFS with Apache Solr How to develop a Kafka messaging system How to develop a Mahout User Recommender System How to stream Logs to HDFS with Apache Flume How to transfer data from MySQL database to Hive, HDFS and HBase with Sqoop How create a Hive table over Apache Solr This book is for: The primary audience is Apache Hadoop developers. Pre-requisite knowledge of Linux and some knowledge of Hadoop is required. If you are a system or application developer interested in learning how to solve practical problems using the Hadoop framework, then this book is made for you. You are expected to be familiar with the Java programming language. Familiarity with Hadoop would be a plus. Learn Big Data from the ground up with this complete and up-to-date resource from leaders in the field Big Data: Concepts, Technology, and Architecture delivers a comprehensive treatment of Big Data tools, terminology, and technology perfectly suited to a wide range of business professionals, academic researchers, and students. Beginning with a full-scope overview of what we mean when we say, “Big Data,” the book moves on to discuss every stage of the lifecycle of Big Data. You’ll learn about the creation of structured, unstructured, and semi-structured data, data storage solutions, traditional database solutions like SQL, data processing, data analytics, machine learning, and data mining. You’ll also discover how specific technologies like Apache Hadoop, SQOOP, and Flume work. Big Data also covers the central topic of big data visualization with Tableau, and you’ll learn how to create scatter plots, histograms, bar, line, and pie charts with that software. Accessibly organized, Big Data includes illuminating case studies throughout the material, showing you step by step how these concepts have been applied in real-world settings. Some of the common challenges facing big data technology and technologists, like data heterogeneity and incompleteness, data volume and velocity, storage limitations, and privacy concerns Relational and non-relational databases, like RDBMS, NoSQL, and NewSQL databases Virtualizing Big Data through encapsulation, partitioning, and isolating, as well as big data server virtualization Apache software, including Hadoop, Cassandra, Avro, Pig, Mahout, Oozie, and Hive The Big Data analytics lifecycle, including business case evaluation, data preparation, extraction, transformation, analysis, and visualization. Perfect for data scientists, data engineers, and database managers, Big Data also belongs on the bookshelves of business intelligence analysts who are required to make decisions based on large volumes of information. Executives and managers who lead teams responsible for keeping or understanding large datasets will also benefit from this book. 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. Integrate Elasticsearch into Hadoop to effectively visualize and analyze your data. About This Book Build production-ready analytics applications by integrating the Hadoop ecosystem with Elasticsearch. Learn complex Elasticsearch queries and develop real-time monitoring. Kibana dashboards to visualize your data Use Elasticsearch and Kibana to search data in Hadoop easily with this comprehensive, step-by-step guide. Who This Book Is For This book is targeted at Java developers with basic knowledge on Hadoop. No prior Elasticsearch experience is expected. You will learn to set up the Elasticsearch-Hadoop environment. Import HDFS data into Elasticsearch with MapReduce jobs. Perform full-text search and aggregations efficiently using Elasticsearch. Search data and create interactive dashboards using Kibana. Check and detect anomalies in streaming data using Storm and Elasticsearch. Inject and classify real-time streaming data into Elasticsearch. Get production-ready for Elasticsearch-Hadoop based projects. Integrate with Hadoop eco-
system such as Pig, Storm, Hive, and Spark. In Detail. The Hadoop ecosystem is a de-facto standard for processing terabytes of data. Lucene-enabled Elasticsearch is becoming an industry standard for its full-text search and aggregation capabilities. Elasticsearch-Hadoop serves as a perfect tool to bridge the worlds of Elasticsearch and Hadoop ecosystem to get best out of both the worlds. Powering with Kibana, this stack makes it a cocoa platform to get surprising insights out of your massive amount of Hadoop ecosystem in a flash. In this book, you’ll learn to use Elasticsearch, Kibana, and Elasticsearch-Hadoop effectively to analyze and understand your HDFS and streaming data. You begin with an in-depth understanding of the Hadoop, Elasticsearch, Marvel, and Kibana setup. Right after this, you will learn to successfully import Hadoop data into Elasticsearch by writing MapReduce job in a real-world example. This is then followed by a comprehensive look at Elasticsearch essentials, such as full-text search analysis, queries, filters, and aggregations; after which you gain an understanding of creating various visualizations and interactive dashboard using Kibana. Classifying your real-world streaming data and identifying trends in it using Storm and Elasticsearch are some of the other topics that we’ll cover. You will also gain an insight about key concepts of Elasticsearch and Elasticsearch-hadoop in distributed mode, advanced configurations along with some common configuration presets you may need for your production deployments. You will have “Go production checklist” and high-level view for cluster administration for post-production. Towards the end, you will adapt Elastic Search with OWL, Style with Kibana, and applications. This book approach has been adopted with real-time examples to help you grasp the concepts easily. “It’s not easy to find such a generous book on big data and databases. Fortunately, this book is the one.” Feng Yu. Computing Reviews. June 28, 2016. This is a book for enterprise architects, database administrators, and developers who need to understand the latest developments in database technologies. It is the book to help you choose the correct database technology at a time when concepts such as Big Data, NoSQL and NewSQL are making what was to be an easy choice into a complex decision with significant implications. The relational database (RDBMS) model completely dominated database technology for over 20 years. Today this “one size fits all” stability has been disrupted by a relatively recent explosion of new database technologies. These paradigm-busting technologies are powering the “Big Data” and “NoSQL” revolutions, as well as forcing fundamental changes in databases across the board. Deciding to use a relational database was once truly a no-brainer, and the various commercial relational databases competed on price, performance, reliability, and ease of use rather than on fundamental architectures. Today we are faced with choices between radically different database technologies. Choosing the right database today is a complex undertaking, with serious economic and technological consequences. Next Generation Databases demystifies today’s new database technologies. The book describes what each technology was designed to solve. It shows how each technology can be used to solve real world application and business problems. Most importantly, this book highlights the architectural differences between technologies that are the critical factors to consider when choosing a database platform for new and upcoming projects. Introduces the new technologies that have revolutionized the database landscape. Describes how each technology can be used to solve specific application or business challenges. Reviews the most popular new wave databases and how they use these new database technologies. There’s a lot of information about big data technologies, but splicing these technologies into an end-to-end enterprise data platform is a daunting task not widely covered. With this practical book, you will learn how to build high performing data warehousing and big data infrastructures. This book is not a technical book aimed at developers, but a hands-on guide for enterprise architects, IT managers, application architects, and data engineers, this book shows you how to overcome the many challenges that emerge during Hadoop projects. You’ll explore the vast landscape of tools available in the Hadoop and big data realm in a thorough technical primer before diving into: Infrastructure: Look at all component layers in a modern data platform, from the server to the data center, to establish a solid foundation for data in your enterprise Platform: Understand aspects of deployment, operation, security, high availability and disaster recovery, along with everything you need to know to integrate your platform with the rest of your enterprise IT? Taking Hadoop to the cloud: Learn the important architectural aspects of running a big data platform in the cloud while maintaining enterprise security and high availability. This book is a complete practical approach for Hadoop lovers. It is mainly aimed at beginners who want to have a hands-on experience with Hadoop and its ecosystem. Its simplicity and step-by-step explanation will help students and other readers in the computer science industry to use this book as a reference manual. The book has been divided into various chapters that cover Hadoop installation, Summary on Hadoop core components, General commands in Hadoop with examples, SQOOP—import & export commands with verification steps, Pig Latin Commands, Analysis using Pig Latin, Pig Script examples, HiveQL Queries and expected outputs and HBase with CRUD operations. In short, this book is a guide for programmers and non-programmers to begin their projects in Hadoop. It is also suitable as a reference manual for projects in Hadoop for students and professionals who are new to the Hadoop Ecosystems, Learn how to write, tune, and port SQL queries and other statements for a Big Data environment, using Impala—the massively parallel processing SQL query engine for Apache Hadoop. The best practices in this practical guide help you design database schemas that not only interoperate with other Hadoop components, and are convenient for administrators to manage and monitor, but also accommodate future expansion in database size and evolution of software capabilities. Written by John Russell, documentation lead for the Cloudera Impala project, this book gets you working with the most recent Impala releases quickly. Ideal for database developers and business analysts, the latest revision covers analytics functions, complex types, incremental statistics, subqueries, and submission to the Apache incubator. Getting Started with Impala includes advice from Cloudera’s development team, as well as insights from its consulting engagements with customers. Learn how Impala integrates with a wide range of Hadoop components to attain high performance and scalability for huge data sets on production clusters. Explore common developer tasks, such as porting code to Impala and optimizing performance. Use tutorials for working with billion-row tables, date- and time-based values, and other techniques. Learn how to transition from rigid schemas to a flexible model that evolves as needs change. Take a deep dive into joins and the roles of statistics. This book presents the outcomes of the 2020 International Conference on Cyber Security Intelligence and Analytics (CSIA 2020), which was dedicated to promoting novel theoretical and applied research advances in the interdisciplinary field of cyber security, particularly those focusing on threat intelligence, analytics, and preventing cyber crime. The conference
provides a forum for presenting and discussing innovative ideas, cutting-edge research findings, and novel techniques, methods, and applications concerning all aspects of cyber security intelligence and analytics. CSIA 2020, which was held in Haikou, China on February 28–29, 2020, built on the previous conference in Wuhu, China (2019), and marks the series’ second successful installment. “In this Introduction to Apache Hive training course, expert author Tom Hanlon will teach you how to create and query large datasets in Hadoop. This course is designed for the absolute beginner, meaning no experience with SQL or Hadoop is required. You will start by learning how to create tables and load data. From there, Tom will teach you how to manipulate tables with HiveQL, including how to insert data into a Hive table using HiveQL. This video tutorial also covers how to create views and partitions and transform data with custom scripts. Finally, you will learn about Hive execution engines, such as Map Reduce, Tez, and Spark. Once you have completed this computer-based training course, you will have learned how to create tables and load data in Hive, execute SQL queries, use user-defined functions, and transform scripts.”—Resource description page.Impala and Hive bring SQL technologies on Hadoop Systems enabling users to run analytics queries against data stored in HDFS and Apache HBase without requiring data movement or transformation. In this work we characterize BigDataBench SQL workloads in Impala as I/O, Communication or Compute intensive. We do detailed profiling and analysis of query execution in Impala to understand performance issues observed in Impala and to identify the existing Join implementation is blocking based. This work implements a non-blocking Join where the reading of right-side table of Join and building of its Hashtable is overlapped with construction of left-side table data. Experimental results show that non-blocking Join implementation improves the execution of Join queries by 9-12%. Next scalability study of Impala is performed to evaluate how well Impala scales out on increasing the number of compute nodes for divergent SQL queries. We observe that the default Inner Join SQL query is not scaling well since Impala by default does a broadcast Join. We change the default Inner Join in Impala to do partitioned/shuffle Join and the results show that it scales linearly. We then evaluate Hive SQL queries running on top of Triple-H – RDMA (Remote Direct Memory Access) based HDFS which is optimized for HDFS-Write. We design new write intensive SQL benchmark queries and the experimental results show that TripleH brings benefit of 45% to write intensive queries and 25% benefit to read intensive query in Hive. In another scheme we evaluate querying of HBase tables in Hive running on top of Triple-H and we see 20-33% benefit for write intensive queries and 15% benefit for read intensive query. From these results we show improvements of SQL queries on Hadoop Systems. 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 how to perform data loading with Hive Use Hive to manipulate data with HiveQL, including how to insert data into a Hive table using HiveQL. This video tutorial also covers how to create views and partitions and transform data with custom scripts. Finally, you will learn about Hive execution engines, such as Map Reduce, Tez, and Spark. Once you have completed this computer-based training course, you will have learned how to create tables and load data in Hive, execute SQL queries, use user-defined functions, and transform scripts.”—Resource description page.Impala and Hive bring SQL technologies on Hadoop Systems enabling users to run analytics queries against data stored in HDFS and Apache HBase without requiring data movement or transformation. In this work we characterize BigDataBench SQL workloads in Impala as I/O, Communication or Compute intensive. We do detailed profiling and analysis of query execution in Impala to understand performance issues observed in Impala and to identify the existing Join implementation is blocking based. This work implements a non-blocking Join where the reading of right-side table of Join and building of its Hashtable is overlapped with construction of left-side table data. Experimental results show that non-blocking Join implementation improves the execution of Join queries by 9-12%. Next scalability study of Impala is performed to evaluate how well Impala scales out on increasing the number of compute nodes for divergent SQL queries. We observe that the default Inner Join SQL query is not scaling well since Impala by default does a broadcast Join. We change the default Inner Join in Impala to do partitioned/shuffle Join and the results show that it scales linearly. We then evaluate Hive SQL queries running on top of Triple-H – RDMA (Remote Direct Memory Access) based HDFS which is optimized for HDFS-Write. We design new write intensive SQL benchmark queries and the experimental results show that TripleH brings benefit of 45% to write intensive queries and 25% benefit to read intensive query in Hive. In another scheme we evaluate qu
beginner, meaning no experience with SQL or Hadoop is required. You will start by learning how to connect to Hive, then jump into learning how to create tables and load data. From there, Tom will teach you how to manipulate tables with HiveQL, including how to insert data into a Hive table using HiveQL. This video tutorial also covers how to create views and partitions and store data with custom scripts. Finally, you will learn about Hive execution engines, such as Map Reduce, Tez, and Spark. Once you have completed this computer-based training course, you will have learned how to create tables and load data in Hive, execute SQL queries, use user-defined functions, and transform scripts.