Hive Sql For Hadoop

Hive allows you to take data in Hadoop, apply a fixed external schema, and query the data with an SQL-like language. With Hive, complex queries can yield simpler, more effectively visualized results. Author Elton Stoneman uses Hive to introduce the core principles of Hive and guides readers through mapping Hadoop and HBase data in Hive, writing complex queries in HiveQL, and running custom

code inside Hive queries using a variety of languages. With this e-book, getting the most out of big data and Hadoop has never been easier. This updated and expanded second edition of Book provides a user-friendly introduction to the subject, Taking a clear structural framework, it guides the reader through the subject's core elements. A flowing writing style combines with the use of illustrations and diagrams throughout the text to ensure the reader

understands even the most complex of concepts. This succinct and enlightening overview is a required reading for all those interested in the subject . We hope you find this book useful in shaping your future career & Business.
Let Hadoop For Dummies help harness the power of your data and rein in the information overload Big data has become big business, and companies and organizations of all sizes are struggling to find ways to retrieve valuable

information from their massive data sets with becoming overwhelmed. Enter Hadoop and this easy-to-understand For Dummies guide. Hadoop For Dummies helps readers understand the value of big data, make a business case for using Hadoop, navigate the Hadoop ecosystem, and build and manage Hadoop applications and clusters. Explains the origins of Hadoop, its economic benefits, and its functionality and practical applications Helps you find your way

around the Hadoop ecosystem, program MapReduce, utilize design patterns, and get your Hadoop cluster up and running quickly and easily Details how to use Hadoop applications for data mining, web analytics and personalization, large-scale text processing, data science, and problem-solving Shows you how to improve the value of your Hadoop cluster, maximize your investment in Hadoop, and avoid common pitfalls when building your Hadoop cluster From

programmers challenged with building and maintaining affordable, scaleable data systems to administrators who must deal with huge volumes of information effectively and efficiently, this how-to has something to help you with Hadoop. Do you want to broaden your Hadoop skill set and take your knowledge to the next level? Do you wish to enhance your knowledge of Hadoop to solve challenging data processing problems? Are your Hadoop jobs, Pig scripts, or

Hive queries not working as fast as you intend? Are you looking to understand the benefits of upgrading Hadoop? If the answer is yes to any of these, this book is for you. It assumes novice-level familiarity with Hadoop.

Get expert guidance on architecting end-to-end data management solutions with Apache Hadoop. While many sources explain how to use various components in the Hadoop ecosystem, this practical book takes you through architectural

considerations necessary to tie those components together into a complete tailored application, based on your particular use case. To reinforce those lessons, the book's second section provides detailed examples of architectures used in some of the most commonly found Hadoop applications. Whether you're designing a new Hadoop application, or planning to integrate Hadoop into your existing data infrastructure, Hadoop Application

Architectures will skillfully guide you through the process. This book covers: Factors to consider when using Hadoop to store and model data Best practices for moving data in and out of the system Data processing frameworks, including MapReduce, Spark, and Hive Common Hadoop processing patterns, such as removing duplicate records and using windowing analytics Giraph, GraphX, and other tools for large graph processing on Hadoop Using workflow orchestration

**and scheduling tools such as Apache
Oozie Near-real-time stream processing
with Apache Storm, Apache Spark
Streaming, and Apache Flume
Architecture examples for clickstream
analysis, fraud detection, and data
warehousing
Hadoop in Action
Hive in a Day
Spark: The Definitive Guide
Hadoop For Dummies
Architecting Modern Data Platforms**

This 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 essential. 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'll learn how to build big data infrastructure both on-premises and in the cloud and successfully architect a modern data platform. Ideal 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 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 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 In Detail In this book, we prepare you for your journey into big data by frstly 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 effcient manner. Toward the end, the book focuses on advanced topics, such as performance, security, and extensions in Hive, which will guide you on exciting adventures on this worthwhile big data journey. By the end of the book, you will be familiar with Hive and able to work effeciently to find solutions to big data problems 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 ...

Filled with 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.

Getting Started with Kudu

SQOOP, PIG, HIVE, HBASE for Beginners

Mastering Hadoop

Learning PySpark

Hive for Newbies

Build data-intensive applications locally and deploy at scale using the combined powers of Python and Spark 2.0 About This Book Learn why and how you can efficiently use Python to process

data and build machine learning models in Apache Spark 2.0 Develop and deploy efficient, scalable real-time Spark solutions Take your understanding of using Spark with Python to the next level with this jump start guide Who This Book Is For If you are a Python developer who wants to learn about the Apache Spark 2.0 ecosystem, this book is for you. A firm understanding of Python is expected to get the best out of the book. Familiarity with Spark would be useful, but is not mandatory. What You Will Learn Learn about Apache Spark and the Spark 2.0 architecture Build and interact with Spark DataFrames using Spark SQL Learn how to solve graph and deep learning problems using GraphFrames and TensorFrames respectively Read, transform, and understand data and use it to train machine learning models Build machine

learning models with MLlib and ML Learn how to submit your applications programmatically using spark-submit Deploy locally built applications to a cluster In Detail Apache Spark is an open source framework for efficient cluster computing with a strong interface for data parallelism and fault tolerance. This book will show you how to leverage the power of Python and put it to use in the Spark ecosystem. You will start by getting a firm understanding of the Spark 2.0 architecture and how to set up a Python environment for Spark. You will get familiar with the modules available in PySpark. You will learn how to abstract data with RDDs and DataFrames and understand the streaming capabilities of PySpark. Also, you will get a thorough overview of machine learning capabilities of PySpark using ML and MLlib,

graph processing using GraphFrames, and polyglot persistence using Blaze. Finally, you will learn how to deploy your applications to the cloud using the spark-submit command. By the end of this book, you will have established a firm understanding of the Spark Python API and how it can be used to build data-intensive applications. Style and approach This book takes a very comprehensive, step-by-step approach so you understand how the Spark ecosystem can be used with Python to develop efficient, scalable solutions. Every chapter is standalone and written in a very easy-to-understand manner, with a focus on both the hows and the whys of each concept. 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

Find the right big data solution for your business ororganization Big data management is one of the major challenges facingbusiness, industry, and not-for-profit organizations. Data setssuch as customer transactions for a mega-retailer, weather patternsmonitored by meteorologists, or social network activity can quicklyoutpace the capacity of traditional data management tools. If youneed to develop or manage big data solutions, you'll

appreciate howthese four experts define, explain, and guide you through this newand often confusing concept. You'll learn what it is, why itmatters, and how to choose and implement solutions that work. Effectively managing big data is an issue of growing importanceto businesses, not-for-profit organizations, government, and ITprofessionals Authors are experts in information management, big data, and avariety of solutions Explains big data in detail and discusses how to select andimplement a solution, security concerns to consider, data storageand 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 helpsyou take charge of big data

solutions for your organization.

Fast data ingestion, serving, and analytics in the Hadoop ecosystem have forced developers and architects to choose solutions using the least common denominator—either fast analytics at the cost of slow data ingestion or fast data ingestion at the cost of slow analytics. There is an answer to this problem. With the Apache Kudu column-oriented data store, you can easily perform fast analytics on fast data. This practical guide shows you how. Begun as an internal project at Cloudera, Kudu is an open source solution compatible with many data processing frameworks in the Hadoop environment. In this book, current and former solutions professionals from Cloudera provide use cases, examples, best practices, and sample code to help you get

up to speed with Kudu. Explore Kudu's high-level design, including how it spreads data across servers Fully administer a Kudu cluster, enable security, and add or remove nodes Learn Kudu's client-side APIs, including how to integrate Apache Impala, Spark, and other frameworks for data manipulation Examine Kudu's schema design, including basic concepts and primitives necessary to make your project successful Explore case studies for using Kudu for real-time IoT analytics, predictive modeling, and in combination with another storage engine Programming Hive
Hadoop in Practice
Performance Characterization and Improvements of SQL-on-hadoop Systems

Hadoop: The Definitive Guide

Expert techniques for architecting end-to-end big data solutions to get valuable insights

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 frstly 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 effcient manner. Toward the end, the book focuses on advanced topics, such as performance, security, and extensions in Hive, which will guide you on exciting adventures on this worthwhile big data journey. By the end of the book, you will be familiar with Hive and able to work effeciently to find solutions to big data problems What you will learn Create and set up the Hive environment Discover how to use Hive's definition language to describe data Discover interesting data by joining and filtering datasets in Hive Transform data by using Hive sorting, ordering, and

functions Aggregate and sample data in different ways Boost Hive query performance and enhance data security in Hive Customize Hive to your needs by using user-defined functions and integrate it with other tools Who this book is for If you are a data analyst, developer, or simply someone who wants to quickly get started with Hive to explore and analyze Big Data in Hadoop, this is the book for you. Since Hive is an SQL-like language, some previous experience with SQL will be useful to get the most out of this book.
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 fulsome 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 how the included concepts have been applied in real-world settings. Some of those concepts include: 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. "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 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 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.

Hive allows you to take data in Hadoop, apply a fixed external schema, and query the data with an SQL-like language. With Hive, complex queries can yield simpler, more

effectively visualized results. Author Elton Stoneman uses Hive to introduce the core principles of Hive and guides readers through mapping Hadoop and HBase data in Hive, writing complex queries in HiveQL, and running custom code inside Hive queries using a variety of languages. With this e-book, getting the most out of big data and Hadoop has never been easier.This updated and expanded second edition of Book provides a user-friendly introduction to the subject, Taking a clear structural framework, it guides the reader through the subject's core elements. A flowing writing style combines with the use of illustrations and diagrams throughout the text to ensure the reader understands even the most complex of concepts. This succinct and enlightening

overview is a required reading for all those interested in the subject . We hope you find this book useful in shaping your future career & Business.

Apache Hive Essentials

Learn Hive in 24 Hours

Modern Big Data Processing with Hadoop

Introduction to Apache Hive

Getting Started with Impala

Apache Hive is the new member in 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 each and 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 Use
Learn how to use, deploy, and maintain Apache Spark with this comprehensive guide, written by the creators of the open-source cluster-computing framework. With an emphasis on improvements and new features in Spark 2.0, authors Bill Chambers and Matei Zaharia break down Spark topics into distinct sections, each with unique goals. You'll explore the basic operations and common

functions of Spark's structured APIs, as well as Structured Streaming, a new high-level API for building end-to-end streaming applications. Developers and system administrators will learn the fundamentals of monitoring, tuning, and debugging Spark, and explore machine learning techniques and scenarios for employing MLlib, Spark's scalable machine-learning library. Get a gentle overview of big data and Spark Learn about DataFrames, SQL, and Datasets—Spark's core APIs—through worked examples Dive into Spark's low-level

APIs, RDDs, and execution of SQL and DataFrames Understand how Spark runs on a cluster Debug, monitor, and tune Spark clusters and applications Learn the power of Structured Streaming, Spark's stream-processing engine Learn how you can apply MLlib to a variety of problems, including classification or recommendation 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. You are expected to

be familiar with the Unix/Linux command-line interface and have some experience with the Java programming language. Familiarity with Hadoop would be a plus.

Harness the power of SQL Server 2017 Integration Services to build your data integration solutions with ease About This Book Acquaint yourself with all the newly introduced features in SQL Server 2017 Integration Services Program and extend your packages to enhance their functionality This detailed, step-by-step guide covers everything

you need to develop efficient data integration and data transformation solutions for your organization Who This Book Is For This book is ideal for software engineers, DW/ETL architects, and ETL developers who need to create a new, or enhance an existing, ETL implementation with SQL Server 2017 Integration Services. This book would also be good for individuals who develop ETL solutions that use SSIS and are keen to learn the new features and capabilities in SSIS 2017. What You Will Learn Understand the key

components of an ETL solution using SQL
Server 2016-2017 Integration Services Design
the architecture of a modern ETL solution
Have a good knowledge of the new
capabilities and features added to Integration
Services Implement ETL solutions using
Integration Services for both on-premises and
Azure data Improve the performance and
scalability of an ETL solution Enhance the ETL
solution using a custom framework Be able to
work on the ETL solution with many other
developers and have common design

paradigms or techniques Effectively use scripting to solve complex data issues In Detail SQL Server Integration Services is a tool that facilitates data extraction, consolidation, and loading options (ETL), SQL Server coding enhancements, data warehousing, and customizations. With the help of the recipes in this book, you'll gain complete hands-on experience of SSIS 2017 as well as the 2016 new features, design and development improvements including SCD, Tuning, and Customizations. At the start, you'll learn to

install and set up SSIS as well other SQL Server resources to make optimal use of this Business Intelligence tools. We'll begin by taking you through the new features in SSIS 2016/2017 and implementing the necessary features to get a modern scalable ETL solution that fits the modern data warehouse. Through the course of chapters, you will learn how to design and build SSIS data warehouses packages using SQL Server Data Tools. Additionally, you'll learn to develop SSIS packages designed to maintain a data

warehouse using the Data Flow and other
control flow tasks. You'll also be demonstrated
many recipes on cleansing data and how to
get the end result after applying different
transformations. Some real-world scenarios
that you might face are also covered and how
to handle various issues that you might face
when designing your packages. At the end of
this book, you'll get to know all the key
concepts to perform data integration and
transformation. You'll have explored on-
premises Big Data integration processes to

create a classic data warehouse, and will know how to extend the toolbox with custom tasks and transforms. Style and approach This cookbook follows a problem-solution approach and tackles all kinds of data integration scenarios by using the capabilities of SQL Server 2016 Integration Services. This book is well supplemented with screenshots, tips, and tricks. Each recipe focuses on a particular task and is written in a very easy-to-follow manner. A Guide for Developers and Administrators Microsoft SQL Server 2012 with Hadoop

Quickly Create and Query Large Datasets with
SQL in Hadoop
Hive for Busies
Cyber Security Intelligence and Analytics
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. Impala and Hive bring SQL technologies on Hadoop

Systems enabling users to run analytics queries aganist 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 the performance of SQL queries. From the analysis we observe that the performance of Inner Join queries can be improved in Impala since 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 Triple-H 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.

Although you don't need a large computing infrastructure to process massive amounts of data with

Apache Hadoop, it can still be difficult to get started. This practical guide shows you how to quickly launch data analysis projects in the cloud by using Amazon Elastic MapReduce (EMR), the hosted Hadoop framework in Amazon Web Services (AWS). Authors Kevin Schmidt and Christopher Phillips demonstrate best practices for using EMR and various AWS and Apache technologies by walking you through the construction of a sample MapReduce log analysis application. Using code samples and example configurations, you'll learn how to assemble the building blocks necessary to solve your biggest data analysis problems. Get an overview of the

AWS and Apache software tools used in large-scale data analysis Go through the process of executing a Job Flow with a simple log analyzer Discover useful MapReduce patterns for filtering and analyzing data sets Use Apache Hive and Pig instead of Java to build a MapReduce Job Flow Learn the basics for using Amazon EMR to run machine learning algorithms Develop a project cost model for using Amazon EMR and other AWS tools
Hadoop Application Architectures
Apache Hive Cookbook
Big Data Processing Made Simple
HBase

Spring Data
Summary Hadoop in Practice, Second Edition provides over 100 tested, instantly useful techniques that will help you conquer big data, using Hadoop. This revised new edition covers changes and new features in the Hadoop core architecture, including MapReduce 2. Brand new chapters cover YARN and integrating Kafka, Impala, and Spark SQL with Hadoop. You'll also get new and updated techniques for Flume, Sqoop, and Mahout, all of which have seen major new versions recently. In short, this is the most practical, up-to-date coverage of Hadoop available anywhere. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub

formats from Manning Publications. About the Book It's always a good time to upgrade your Hadoop skills! Hadoop in Practice, Second Edition provides a collection of 104 tested, instantly useful techniques for analyzing real-time streams, moving data securely, machine learning, managing large-scale clusters, and taming big data using Hadoop. This completely revised edition covers changes and new features in Hadoop core, including MapReduce 2 and YARN. You'll pick up hands-on best practices for integrating Spark, Kafka, and Impala with Hadoop, and get new and updated techniques for the latest versions of Flume, Sqoop, and Mahout. In short, this is the most practical, up-to-date coverage of Hadoop

available. Readers need to know a programming language like Java and have basic familiarity with Hadoop. What's Inside Thoroughly updated for Hadoop 2 How to write YARN applications Integrate real-time technologies like Storm, Impala, and Spark Predictive analytics using Mahout and RR Readers need to know a programming language like Java and have basic familiarity with Hadoop. About the Author Alex Holmes works on tough big-data problems. He is a software engineer, author, speaker, and blogger specializing in large-scale Hadoop projects. Table of Contents PART 1 BACKGROUND AND FUNDAMENTALS Hadoop in a heartbeat Introduction to YARN PART 2 DATA LOGISTICS Data serialization—working with

text and beyond Organizing and optimizing data in HDFS Moving data into and out of Hadoop PART 3 BIG DATA PATTERNS Applying MapReduce patterns to big data Utilizing data structures and algorithms at scale Tuning, debugging, and testing PART 4 BEYOND MAPREDUCE SQL on Hadoop Writing a YARN application

Hadoop in Action teaches readers how to use Hadoop and write MapReduce programs. The intended readers are programmers, architects, and project managers who have to process large amounts of data offline. Hadoop in Action will lead the reader from obtaining a copy of Hadoop to setting it up in a cluster and writing data analytic programs. The book

begins by making the basic idea of Hadoop and MapReduce easier to grasp by applying the default Hadoop installation to a few easy-to-follow tasks, such as analyzing changes in word frequency across a body of documents. The book continues through the basic concepts of MapReduce applications developed using Hadoop, including a close look at framework components, use of Hadoop for a variety of data analysis tasks, and numerous examples of Hadoop in action. Hadoop in Action will explain how to use Hadoop and present design patterns and practices of programming MapReduce. MapReduce is a complex idea both conceptually and in its implementation, and Hadoop users are challenged to

learn all the knobs and levers for running Hadoop. This book takes you beyond the mechanics of running Hadoop, teaching you to write meaningful programs in a MapReduce framework. This book assumes the reader will have a basic familiarity with Java, as most code examples will be written in Java. Familiarity with basic statistical concepts (e.g. histogram, correlation) will help the reader appreciate the more advanced data processing examples. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book. 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 Describes the features and functions of Apache Hive, the data infrastructure for Hadoop.
HADOOP
Programming Elastic MapReduce

**The Definitive Guide
Using AWS Services to Build an End-to-End
Application
Big Data**

This book presents the outcomes of the 2019
International Conference on Cyber Security
Intelligence and Analytics (CSIA2019), an
international conference dedicated to
promoting novel theoretical and applied
research advances in the interdisciplinary field
of cyber security, particularly focusing on
threat intelligence, analytics, and countering

cyber crime. The conference provides a forum for presenting and discussing innovative ideas, cutting-edge research findings, and novel techniques, methods and applications on all aspects of Cyber Security Intelligence and Analytics.

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 students and professionals who are new to the Hadoop Ecosystems. 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 Delve into the big data technologies required to meet modern day business strategies A highly practical guide to

implementing enterprise data lakes with lots of examples and real-world 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, along with several real-world use-cases. It also shows you how other peripheral components can be added to the lake to make it more efficient. 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 and approach The book takes a pragmatic approach, showing ways to leverage big data technologies and lambda architecture to build an enterprise-level data lake.

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
SQL Server 2017 Integration Services
Cookbook
Instant Apache Hive Essentials How-to
Hadoop Practice Guide

A Guide to Hadoop's Data Warehouse System
Concepts, Technology, and Architecture
Dive into the world of SQL on Hadoop and get the most out of your Hive data warehouses. This book is your go-to resource for using Hive: authors Scott Shaw, Ankur Gupta, David Kjerrumgaard, and Andreas Francois Vermeulen take you through learning HiveQL, the SQL-like language specific to Hive, to analyze, export, and massage the data stored across your Hadoop environment. From deploying Hive on your hardware or virtual machine and setting up its initial configuration to learning how Hive interacts with Hadoop, MapReduce, Tez and other big

data technologies, Practical Hive gives you a detailed treatment of the software. In addition, this book discusses the value of open source software, Hive performance tuning, and how to leverage semi-structured and unstructured data. What You Will Learn Install and configure Hive for new and existing datasets Perform DDL operations Execute efficient DML operations Use tables, partitions, buckets, and user-defined functions Discover performance tuning tips and Hive best practices Who This Book Is For Developers, companies, and professionals who deal with large amounts of data and could use software that can efficiently manage large volumes of input. It is assumed that readers have the ability to work with

SQL.

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 the 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.

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. You can choose several data access frameworks when building Java enterprise applications that work with relational databases. But what about big data? This hands-on introduction shows you how Spring Data makes it relatively easy to build applications across a wide range of new data access technologies such as NoSQL and Hadoop. Through several sample projects, you'll learn how Spring Data provides a consistent programming model that retains NoSQL-specific features and capabilities, and helps you develop Hadoop applications across a wide range of use-cases such as data analysis, event stream processing, and

workflow. You'll also discover the features Spring Data adds to Spring's existing JPA and JDBC support for writing RDBMS-based data access layers. Learn about Spring's template helper classes to simplify the use of database-specific functionality Explore Spring Data's repository abstraction and advanced query functionality Use Spring Data with Redis (key/value store), HBase (column-family), MongoDB (document database), and Neo4j (graph database) Discover the GemFire distributed data grid solution Export Spring Data JPA-managed entities to the Web as RESTful web services Simplify the development of HBase applications, using a lightweight object-mapping framework Build example big-data pipelines with

Spring Batch and Spring Integration
Practical Hive
A Guide to Enterprise Hadoop at Scale
Programming Pig
Hadoop Operations
Essential techniques to help you process, and get unique insights from, big data, 2nd Edition
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 administers to manage and monitor, but also accommodate future expansion in data size and evolution of software capabilities. Written by John Russell, documentation lead for the Cloudera Impala project, this book gets you working with the most recent Impala releases quickly. Ideal for database developers and business analysts, the latest revision covers analytics functions, complex types, incremental statistics, subqueries, and submission to the Apache incubator. Getting Started with Impala includes advice from Cloudera's development team, as well as insights from its consulting engagements with customers. Learn how Impala integrates with a wide range of Hadoop

components Attain high performance and scalability for huge data sets on production clusters Explore common developer tasks, such as porting code to Impala and optimizing performance Use tutorials for working with billion-row tables, date- and time-based values, and other techniques Learn how to transition from rigid schemas to a flexible model that evolves as needs change Take a deep dive into joins and the roles of statistics
Perform Fast Analytics on Fast Data
Data Lake for Enterprises
Interactive SQL for Apache Hadoop
Big Data For Dummies
Learning Hadoop 2