

Apache Hive Essentials

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.

This book provides an introduction to Apache HTTP Server - a free, open-source web server. Apache is the most popular web server software on the Internet; it is estimated

Acces PDF Apache Hive Essentials

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

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.

In 2019, America is bordering on financial collapse after engaging in a third conflict in

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

This book takes you on a fantastic journey to discover the attributes of big data using Apache Hive. Key Features Grasp the skills needed to write efficient Hive queries to analyze the Big Data Discover how Hive can coexist and work with other tools within the Hadoop ecosystem Uses practical, example-oriented scenarios to cover all the newly released features of Apache Hive 2.3.3 Book Description In this book, we prepare you for your journey into big data by firstly introducing you to backgrounds in the big data domain, alongwith the process of setting up and getting familiar with your

Acces PDF Apache Hive Essentials

Hive working environment. Next, the book guides you through discovering and transforming the values of big data with the help of examples. It also hones your skills in using the Hive language in an efficient manner. Toward the end, the book focuses on advanced topics, such as performance, security, and extensions in Hive, which will guide you on exciting adventures on this worthwhile big data journey. By the end of the book, you will be familiar with Hive and able to work effeciently to find solutions to big data problems

What you will learn

- Create and set up the Hive environment
- Discover how to use Hive's definition language to describe data
- Discover interesting data by joining and filtering datasets in Hive
- Transform data by using Hive sorting, ordering, and functions
- Aggregate and sample data in different ways
- Boost Hive query performance and enhance data security in Hive
- Customize Hive to your needs by using user-defined functions and integrate it with other tools

Who this book is for

If you are a data analyst, developer, or simply someone who wants to quickly get started with Hive to explore and analyze Big Data in Hadoop, this is the book for you. Since Hive is an SQL-like language, some previous experience with SQL will be useful to get the most out of this book.

Get Started Fast with Apache Hadoop® 2, YARN, and Today's Hadoop Ecosystem

With Hadoop 2.x and YARN, Hadoop moves beyond MapReduce to become practical for virtually any type of data processing. Hadoop 2.x and the Data Lake concept represent a radical shift away from conventional approaches to data usage and

storage. Hadoop 2.x installations offer unmatched scalability and breakthrough extensibility that supports new and existing Big Data analytics processing methods and models. Hadoop® 2 Quick-Start Guide is the first easy, accessible guide to Apache Hadoop 2.x, YARN, and the modern Hadoop ecosystem. Building on his unsurpassed experience teaching Hadoop and Big Data, author Douglas Eadline covers all the basics you need to know to install and use Hadoop 2 on personal computers or servers, and to navigate the powerful technologies that complement it. Eadline concisely introduces and explains every key Hadoop 2 concept, tool, and service, illustrating each with a simple “beginning-to-end” example and identifying trustworthy, up-to-date resources for learning more. This guide is ideal if you want to learn about Hadoop 2 without getting mired in technical details. Douglas Eadline will bring you up to speed quickly, whether you’re a user, admin, devops specialist, programmer, architect, analyst, or data scientist. Coverage Includes Understanding what Hadoop 2 and YARN do, and how they improve on Hadoop 1 with MapReduce Understanding Hadoop-based Data Lakes versus RDBMS Data Warehouses Installing Hadoop 2 and core services on Linux machines, virtualized sandboxes, or clusters Exploring the Hadoop Distributed File System (HDFS) Understanding the essentials of MapReduce and YARN application programming Simplifying programming and data movement with Apache Pig, Hive, Sqoop, Flume, Oozie, and HBase Observing application progress, controlling jobs, and managing workflows Managing Hadoop efficiently with Apache

Ambari—including recipes for HDFS to NFSv3 gateway, HDFS snapshots, and YARN configuration Learning basic Hadoop 2 troubleshooting, and installing Apache Hue and Apache Spark

Lots of Kids write letters to Santa, but those delivered to the North Pole are answered by a group of dedicated volunteers who call themselves The Elves. *Blame It On Mistletoe* – Abby Baxter has spent the year since her husband’s death trying to hold on. When she discovers her son is missing, her entire world trembles— until her husband’s best friend appears at her door. Secretly in love with Abby for years, Frank Machado is determined to see JD back in his mother’s arms. Sparks fly, hearts warm, love—and Christmas—are in the air. Should they Blame it on Mistletoe? *If Only In My Dreams* – Jilted in North Pole, Alaska, café owner Amelia Beckett’s bad man-karma has struck again! She wants out of this cutesy town—until a snarling, injured fox in her backyard sends her running to strong, silent neighbor and café regular, Wes Curtis. Wes moved to Alaska after his wife died, not expecting he’d need to brush up admittedly rusty dating skills. But moonlit nights spent helping beautiful, skittish Amelia and the fox relax and heal make him determined to convince Amelia she belongs in Alaska—with him *What Child is This?* - Hope Grayson’s six-year-old daughter clearly wants a daddy for Christmas. Eli Thompson has never forgotten Hope, realizing just how much he’s missed her. When he unexpectedly shows up to help in the clinic, Hope is stunned. She wants to protect her daughter and her heart, but is it possible Eli is the

perfect Christmas present for them both?

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.

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

Acces PDF Apache Hive Essentials

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.

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 Data Science and Big Data Analytics is about harnessing the power of data for new insights. The book covers the breadth of activities and methods and tools that Data Scientists use. The content focuses on concepts, principles and practical applications that are applicable to any industry and technology environment, and the learning is supported and explained with examples that you can replicate using open-source software. This book will help you: Become a contributor on a data science team Deploy

Access PDF Apache Hive Essentials

a structured lifecycle approach to data analytics problems Apply appropriate analytic techniques and tools to analyzing big data Learn how to tell a compelling story with data to drive business action Prepare for EMC Proven Professional Data Science Certification Corresponding data sets are available from the book's page at Wiley which you can find on the Wiley site by searching for the ISBN 9781118876138. Get started discovering, analyzing, visualizing, and presenting data in a meaningful way today!

Perform fast interactive analytics against different data sources using the Trino high-performance distributed SQL query engine. With this practical guide, you'll learn how to conduct analytics on data where it lives, whether it's Hive, Cassandra, a relational database, or a proprietary data store. Analysts, software engineers, and production engineers will learn how to manage, use, and even develop with Trino. Initially developed by Facebook, open source Trino is now used by Netflix, Airbnb, LinkedIn, Twitter, Uber, and many other companies. Matt Fuller, Manfred Moser, and Martin Traverso show you how a single Trino query can combine data from multiple sources to allow for analytics across your entire organization. Get started: Explore Trino's use cases and learn about tools that will help you connect to Trino and query data Go deeper: Learn Trino's internal workings, including how to connect to and query data sources with support for SQL statements, operators, functions, and more Put Trino in production: Secure Trino, monitor workloads, tune queries, and connect more

applications; learn how other organizations apply Trino
Big Data Analytics with R and Hadoop is a tutorial style book that focuses on all the powerful big data tasks that can be achieved by integrating R and Hadoop. This book is ideal for R developers who are looking for a way to perform big data analytics with Hadoop. This book is also aimed at those who know Hadoop and want to build some intelligent applications over Big data with R packages. It would be helpful if readers have basic knowledge of R.

A handy reference guide for data analysts and data scientists to help to obtain value from big data analytics using Spark on Hadoop clusters
About This Book This book is based on the latest 2.0 version of Apache Spark and 2.7 version of Hadoop integrated with most commonly used tools. Learn all Spark stack components including latest topics such as DataFrames, DataSets, GraphFrames, Structured Streaming, DataFrame based ML Pipelines and SparkR. Integrations with frameworks such as HDFS, YARN and tools such as Jupyter, Zeppelin, NiFi, Mahout, HBase Spark Connector, GraphFrames, H2O and Hivemall. Who This Book Is For Though this book is primarily aimed at data analysts and data scientists, it will also help architects, programmers, and practitioners. Knowledge of either Spark or Hadoop would be beneficial. It is assumed that you have basic programming background in Scala, Python, SQL, or R programming with basic Linux experience. Working experience within big data environments is not mandatory. What You Will Learn Find out and

implement the tools and techniques of big data analytics using Spark on Hadoop clusters with wide variety of tools used with Spark and Hadoop Understand all the Hadoop and Spark ecosystem components Get to know all the Spark components: Spark Core, Spark SQL, DataFrames, DataSets, Conventional and Structured Streaming, MLLib, ML Pipelines and Graphx See batch and real-time data analytics using Spark Core, Spark SQL, and Conventional and Structured Streaming Get to grips with data science and machine learning using MLLib, ML Pipelines, H2O, Hivemall, Graphx, SparkR and Hivemall. In Detail Big Data Analytics book aims at providing the fundamentals of Apache Spark and Hadoop. All Spark components – Spark Core, Spark SQL, DataFrames, Data sets, Conventional Streaming, Structured Streaming, MLLib, Graphx and Hadoop core components – HDFS, MapReduce and Yarn are explored in greater depth with implementation examples on Spark + Hadoop clusters. It is moving away from MapReduce to Spark. So, advantages of Spark over MapReduce are explained at great depth to reap benefits of in-memory speeds. DataFrames API, Data Sources API and new Data set API are explained for building Big Data analytical applications. Real-time data analytics using Spark Streaming with Apache Kafka and HBase is covered to help building streaming applications. New Structured streaming concept is explained with an IOT (Internet of Things) use case. Machine learning techniques are covered using MLLib, ML Pipelines and SparkR and Graph Analytics are covered with GraphX and GraphFrames components of Spark. Readers will also

Acces PDF Apache Hive Essentials

get an opportunity to get started with web based notebooks such as Jupyter, Apache Zeppelin and data flow tool Apache NiFi to analyze and visualize data. Style and approach This step-by-step pragmatic guide will make life easy no matter what your level of experience. You will deep dive into Apache Spark on Hadoop clusters through ample exciting real-life examples. Practical tutorial explains data science in simple terms to help programmers and data analysts get started with Data Science 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.

Apache Hive EssentialsPackt Publishing Ltd

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

Access PDF Apache Hive Essentials

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

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.

The Complete Guide to Data Science with Hadoop—For Technical Professionals, Businesspeople, and Students Demand is soaring for professionals who can solve real data science problems with Hadoop and Spark. Practical Data Science with Hadoop® and Spark is your complete guide to doing just that. Drawing on immense experience with Hadoop and big data, three leading experts bring together everything you need: high-level concepts, deep-dive techniques, real-world use cases, practical applications, and hands-on tutorials. The authors introduce the essentials of data science and the modern Hadoop ecosystem, explaining how Hadoop and Spark have evolved into an effective platform for solving data science problems at scale. In addition to comprehensive application coverage, the authors also provide useful guidance on the important steps of data ingestion, data munging, and visualization. Once the groundwork is in place, the authors focus on specific applications, including machine learning,

Access PDF Apache Hive Essentials

predictive modeling for sentiment analysis, clustering for document analysis, anomaly detection, and natural language processing (NLP). This guide provides a strong technical foundation for those who want to do practical data science, and also presents business-driven guidance on how to apply Hadoop and Spark to optimize ROI of data science initiatives. Learn What data science is, how it has evolved, and how to plan a data science career How data volume, variety, and velocity shape data science use cases Hadoop and its ecosystem, including HDFS, MapReduce, YARN, and Spark Data importation with Hive and Spark Data quality, preprocessing, preparation, and modeling Visualization: surfacing insights from huge data sets Machine learning: classification, regression, clustering, and anomaly detection Algorithms and Hadoop tools for predictive modeling Cluster analysis and similarity functions Large-scale anomaly detection NLP: applying data science to human language Data is bigger, arrives faster, and comes in a variety of formats—and it all needs to be processed at scale for analytics or machine learning. But how can you process such varied workloads efficiently? Enter Apache Spark. Updated to include Spark 3.0, this second edition shows data engineers and data scientists why structure and unification in Spark matters. Specifically, this book explains how to perform simple and complex data analytics and employ machine learning algorithms. Through step-by-step walk-throughs, code snippets, and notebooks, you'll be able to: Learn Python, SQL, Scala, or Java high-level Structured APIs Understand Spark operations and SQL Engine Inspect, tune, and debug Spark operations with Spark configurations and Spark UI Connect to data sources: JSON, Parquet, CSV, Avro, ORC, Hive, S3, or Kafka Perform analytics on batch and streaming data using Structured Streaming Build reliable data pipelines with open source Delta Lake and Spark Develop machine learning

Acces PDF Apache Hive Essentials

pipelines with MLlib and productionize models using MLflow

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

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.

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

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.

Use this practical guide to successfully handle the challenges encountered when designing an enterprise data lake and learn industry best practices to resolve issues. When designing an enterprise data lake you often hit a roadblock when you must leave the comfort of the relational world and learn the nuances of handling non-relational data. Starting from sourcing data into the Hadoop ecosystem, you will go through stages that can bring up tough questions such as data processing, data querying, and security. Concepts such as change data capture and data streaming are covered. The book takes an end-to-end solution approach in a data lake

Acces PDF Apache Hive Essentials

environment that includes data security, high availability, data processing, data streaming, and more. Each chapter includes application of a concept, code snippets, and use case demonstrations to provide you with a practical approach. You will learn the concept, scope, application, and starting point. What You'll Learn Get to know data lake architecture and design principles Implement data capture and streaming strategies Implement data processing strategies in Hadoop Understand the data lake security framework and availability model Who This Book Is For Big data architects and solution architects

Integrating data from multiple sources is essential in the age of big data, but it can be a challenging and time-consuming task. This handy cookbook provides dozens of ready-to-use recipes for using Apache Sqoop, the command-line interface application that optimizes data transfers between relational databases and Hadoop. Sqoop is both powerful and bewildering, but with this cookbook's problem-solution-discussion format, you'll quickly learn how to deploy and then apply Sqoop in your environment. The authors provide MySQL, Oracle, and PostgreSQL database examples on GitHub that you can easily adapt for SQL Server, Netezza, Teradata, or other relational systems. Transfer data from a single database table into your Hadoop ecosystem Keep table data and Hadoop in sync by importing data incrementally Import data from more than one database table Customize transferred data by calling various database functions Export generated, processed, or backed-up data from Hadoop to your database Run Sqoop within Oozie, Hadoop's specialized workflow scheduler Load data into Hadoop's data warehouse (Hive) or database (HBase) Handle installation, connection, and syntax issues common to specific database vendors

This book is intended for developers and Big Data engineers who want to know all about

HBase at a hands-on level. For in-depth understanding, it would be helpful to have a bit of familiarity with HDFS and MapReduce programming concepts with no prior experience with HBase or similar technologies. This book is also for Big Data enthusiasts and database developers who have worked with other NoSQL databases and now want to explore HBase as another futuristic, scalable database solution in the Big Data space.

A revelatory exploration of the hottest trend in technology and the dramatic impact it will have on the economy, science, and society at large. Which paint color is most likely to tell you that a used car is in good shape? How can officials identify the most dangerous New York City manholes before they explode? And how did Google searches predict the spread of the H1N1 flu outbreak? The key to answering these questions, and many more, is big data. “Big data” refers to our burgeoning ability to crunch vast collections of information, analyze it instantly, and draw sometimes profoundly surprising conclusions from it. This emerging science can translate myriad phenomena—from the price of airline tickets to the text of millions of books—into searchable form, and uses our increasing computing power to unearth epiphanies that we never could have seen before. A revolution on par with the Internet or perhaps even the printing press, big data will change the way we think about business, health, politics, education, and innovation in the years to come. It also poses fresh threats, from the inevitable end of privacy as we know it to the prospect of being penalized for things we haven’t even done yet, based on big data’s ability to predict our future behavior. In this brilliantly clear, often surprising work, two leading experts explain what big data is, how it will change our lives, and what we can do to protect ourselves from its hazards. Big Data is the first big book about the next big thing. www.big-data-book.com

Acces PDF Apache Hive Essentials

A fast paced guide that will help you learn about Apache Hadoop 3 and its ecosystem Key Features Set up, configure and get started with Hadoop to get useful insights from large data sets Work with the different components of Hadoop such as MapReduce, HDFS and YARN Learn about the new features introduced in Hadoop 3 Book Description Apache Hadoop is a widely used distributed data platform. It enables large datasets to be efficiently processed instead of using one large computer to store and process the data. This book will get you started with the Hadoop ecosystem, and introduce you to the main technical topics, including MapReduce, YARN, and HDFS. The book begins with an overview of big data and Apache Hadoop. Then, you will set up a pseudo Hadoop development environment and a multi-node enterprise Hadoop cluster. You will see how the parallel programming paradigm, such as MapReduce, can solve many complex data processing problems. The book also covers the important aspects of the big data software development lifecycle, including quality assurance and control, performance, administration, and monitoring. You will then learn about the Hadoop ecosystem, and tools such as Kafka, Sqoop, Flume, Pig, Hive, and HBase. Finally, you will look at advanced topics, including real time streaming using Apache Storm, and data analytics using Apache Spark. By the end of the book, you will be well versed with different configurations of the Hadoop 3 cluster. What you will learn Store and analyze data at scale using HDFS, MapReduce and YARN Install and configure Hadoop 3 in different modes Use Yarn effectively to run different applications on Hadoop based platform Understand and monitor how Hadoop cluster is managed Consume streaming data using Storm, and then analyze it using Spark Explore Apache Hadoop ecosystem components, such as Flume, Sqoop, HBase, Hive, and Kafka Who this book is for Aspiring Big Data professionals who want

Acces PDF Apache Hive Essentials

to learn the essentials of Hadoop 3 will find this book to be useful. Existing Hadoop users who want to get up to speed with the new features introduced in Hadoop 3 will also benefit from this book. Having knowledge of Java programming will be an added advantage.

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

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 firstly introducing you to backgrounds in the big data domain, alongwith the process of setting up and getting familiar with your Hive working environment. Next, the book guides you through discovering and transforming the values of big data with the help of examples. It also hones your skills in using the Hive language in an efficient manner. Toward the end, the book focuses on advanced topics, such as performance, security, and extensions

in Hive, which will guide you on exciting adventures on this worthwhile big data journey. By the end of the book, you will be familiar with Hive and able to work effeciently to find solutions to big data problems 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 ...

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

Acces PDF Apache Hive Essentials

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.

Get ready to unlock the power of your data. With the fourth edition of this comprehensive guide, you'll learn how to build and maintain reliable, scalable, distributed systems with Apache Hadoop. This book is ideal for programmers looking to analyze datasets of any size, and for administrators who want to set up and run Hadoop clusters. Using Hadoop 2 exclusively, author Tom White presents new chapters on YARN and several Hadoop-related projects such as Parquet, Flume, Crunch, and Spark. You'll learn about recent changes to Hadoop, and explore new case studies on Hadoop's role in healthcare systems and genomics data processing. Learn fundamental components such as MapReduce, HDFS, and YARN Explore MapReduce in depth, including steps for developing applications with it Set up and maintain a Hadoop cluster running HDFS and MapReduce on YARN Learn two data formats: Avro for data serialization and Parquet for nested data Use data ingestion tools such as Flume (for streaming data) and Sqoop (for bulk data transfer) Understand how high-level data processing tools like Pig, Hive, Crunch, and Spark work with Hadoop Learn the HBase

Acces PDF Apache Hive Essentials

distributed database and the ZooKeeper distributed configuration service

Every enterprise application creates data, whether it's log messages, metrics, user activity, outgoing messages, or something else. And how to move all of this data becomes nearly as important as the data itself. If you're an application architect, developer, or production engineer new to Apache Kafka, this practical guide shows you how to use this open source streaming platform to handle real-time data feeds. Engineers from Confluent and LinkedIn who are responsible for developing Kafka explain how to deploy production Kafka clusters, write reliable event-driven microservices, and build scalable stream-processing applications with this platform. Through detailed examples, you'll learn Kafka's design principles, reliability guarantees, key APIs, and architecture details, including the replication protocol, the controller, and the storage layer. Understand publish-subscribe messaging and how it fits in the big data ecosystem. Explore Kafka producers and consumers for writing and reading messages

Understand Kafka patterns and use-case requirements to ensure reliable data delivery Get best practices for building data pipelines and applications with Kafka Manage Kafka in production, and learn to perform monitoring, tuning, and maintenance tasks Learn the most critical metrics among Kafka's operational measurements Explore how Kafka's stream delivery capabilities make it a perfect source for stream processing systems

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 IBM's Platform of Hadoop framework. IBM Infosphere BigInsight has the highest amount of tutorial matter available free of cost on Internet which makes it easy to

Acces PDF Apache Hive Essentials

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 Jaql, 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.

Unleash the power of Apache Oozie to create and manage your big data and machine learning pipelines in one go About This Book Teaches you everything you need to know to get started with Apache Oozie from scratch and manage your data pipelines effortlessly Learn to write data ingestion workflows with the help of real-life examples from the author's own personal experience Embed Spark jobs to run your machine learning models on top of Hadoop Who This Book Is For If you are an expert Hadoop user who wants to use Apache Oozie to handle workflows efficiently, this book is for you. This book will be handy to anyone who is familiar with the basics of Hadoop and wants to automate data and machine learning pipelines. What You Will Learn Install and configure Oozie from source code on your Hadoop cluster Dive into the world of Oozie with Java MapReduce jobs Schedule Hive ETL and data ingestion jobs Import data from a database through Sqoop jobs in HDFS Create and process data pipelines with Pig, hive scripts as per business requirements. Run machine learning Spark jobs on Hadoop Create quick Oozie jobs using Hue Make the most of Oozie's security capabilities by configuring Oozie's security In Detail As more and more organizations are discovering the use of big data analytics, interest in platforms that provide storage, computation, and analytic capabilities is booming exponentially. This calls for data management. Hadoop caters to this

Access PDF Apache Hive Essentials

need. Oozie fulfills this necessity for a scheduler for a Hadoop job by acting as a cron to better analyze data. Apache Oozie Essentials starts off with the basics right from installing and configuring Oozie from source code on your Hadoop cluster to managing your complex clusters. You will learn how to create data ingestion and machine learning workflows. This book is sprinkled with the examples and exercises to help you take your big data learning to the next level. You will discover how to write workflows to run your MapReduce, Pig ,Hive, and Sqoop scripts and schedule them to run at a specific time or for a specific business requirement using a coordinator. This book has engaging real-life exercises and examples to get you in the thick of things. Lastly, you'll get a grip of how to embed Spark jobs, which can be used to run your machine learning models on Hadoop. By the end of the book, you will have a good knowledge of Apache Oozie. You will be capable of using Oozie to handle large Hadoop workflows and even improve the availability of your Hadoop environment. Style and approach This book is a hands-on guide that explains Oozie using real-world examples. Each chapter is blended beautifully with fundamental concepts sprinkled in-between case study solution algorithms and topped off with self-learning exercises.

Can we add value to the current Apache Hive decision-making process (largely qualitative) by incorporating uncertainty modeling (more quantitative)? Apache Hive in management -Strategic planning How will the Apache Hive team and the organization measure complete success of Apache Hive? Will Apache Hive deliverables need to be tested and, if so, by whom? Who will be responsible for deciding whether Apache Hive goes ahead or not after the initial investigations? This premium Apache Hive self-assessment will make you the credible Apache Hive domain auditor by revealing just what you need to know to be fluent and ready for

Access PDF Apache Hive Essentials

any Apache Hive challenge. How do I reduce the effort in the Apache Hive work to be done to get problems solved? How can I ensure that plans of action include every Apache Hive task and that every Apache Hive outcome is in place? How will I save time investigating strategic and tactical options and ensuring Apache Hive costs are low? How can I deliver tailored Apache Hive advice instantly with structured going-forward plans? There's no better guide through these mind-expanding questions than acclaimed best-selling author Gerard Blokdyk. Blokdyk ensures all Apache Hive essentials are covered, from every angle: the Apache Hive self-assessment shows succinctly and clearly that what needs to be clarified to organize the required activities and processes so that Apache Hive outcomes are achieved. Contains extensive criteria grounded in past and current successful projects and activities by experienced Apache Hive practitioners. Their mastery, combined with the easy elegance of the self-assessment, provides its superior value to you in knowing how to ensure the outcome of any efforts in Apache Hive are maximized with professional results. Your purchase includes access details to the Apache Hive self-assessment dashboard download which gives you your dynamically prioritized projects-ready tool and shows you exactly what to do next. Your exclusive instant access details can be found in your book. You will receive the following contents with New and Updated specific criteria: - The latest quick edition of the book in PDF - The latest complete edition of the book in PDF, which criteria correspond to the criteria in... - The Self-Assessment Excel Dashboard, and... - Example pre-filled Self-Assessment Excel Dashboard to get familiar with results generation ...plus an extra, special, resource that helps you with project managing. INCLUDES LIFETIME SELF ASSESSMENT UPDATES Every self assessment comes with Lifetime Updates and Lifetime Free Updated Books. Lifetime Updates

Acces PDF Apache Hive Essentials

is an industry-first feature which allows you to receive verified self assessment updates, ensuring you always have the most accurate information at your fingertips.

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

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

Access PDF Apache Hive Essentials

(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

Integrate, deploy, rapidly configure, and successfully manage your own big data-intensive clusters in the cloud using OpenStack Sahara About This Book A fast paced guide to help you utilize the benefits of Sahara in OpenStack to meet the Big Data world of Hadoop. A step by step approach to simplify the complexity of Hadoop configuration, deployment and maintenance. Who This Book Is For This book targets data scientists, cloud developers and Devops Engineers who would like to become proficient with OpenStack Sahara. Ideally, this book is well suitable for readers who are familiar with databases, Hadoop and Spark solutions. Additionally, a basic prior knowledge of OpenStack is expected. The readers should also be familiar with different Linux boxes, distributions and virtualization technology. What You Will Learn Integrate and Install Sahara with OpenStack environment Learn Sahara architecture under the hood Rapidly configure and scale Hadoop clusters on top of OpenStack Explore the Sahara REST API to create, deploy and manage a Hadoop cluster Learn the Elastic Processing Data (EDP) facility to execute jobs in clusters from Sahara Cover other Hadoop stable plugins existing supported by Sahara Discover different features provided by Sahara for

Acces PDF Apache Hive Essentials

Hadoop provisioning and deployment Learn how to troubleshoot OpenStack Sahara issues In Detail The Sahara project is a module that aims to simplify the building of data processing capabilities on OpenStack. The goal of this book is to provide a focused, fast paced guide to installing, configuring, and getting started with integrating Hadoop with OpenStack, using Sahara. The book should explain to users how to deploy their data-intensive Hadoop and Spark clusters on top of OpenStack. It will also cover how to use the Sahara REST API, how to develop applications for Elastic Data Processing on Openstack, and setting up hadoop or spark clusters on Openstack. Style and approach This book takes a step by step approach teaching how to integrate, deploy and manage data using OpenStack Sahara. It will teach how the OpenStack Sahara is beneficial by simplifying the complexity of Hadoop configuration, deployment and maintenance.

[Copyright: d17c94d67897b3cff7f841accbc58230](https://www.pdfdrive.com/apache-hive-essentials-p123456789.html)