Hadoop Operations And Cluster Management Cookbook

Thank you very much for downloading **Hadoop Operations And Cluster Management Cookbook**.Most likely you have knowledge that, people have look numerous times for their favorite books following this Hadoop Operations And Cluster Management Cookbook, but stop in the works in harmful downloads.

Rather than enjoying a good PDF in the manner of a mug of coffee in the afternoon, on the other hand they juggled in the manner of some harmful virus inside their computer. **Hadoop Operations And Cluster Management Cookbook** is approachable in our digital library an online entry to it is set as public in view of that you can download it instantly. Our digital library saves in complex countries, allowing you to acquire the most less latency times to download any of our books past this one. Merely said, the Hadoop Operations And Cluster Management Cookbook is universally compatible taking into account any devices to read.

Spring Recipes Daniel Rubio 2014-11-14 Spring Recipes: A Problem-Solution Approach, Third Edition builds upon the best-selling success of the previous editions and focuses on the latest Spring Framework features for building enterprise Java applications. This book provides code recipes for the following, found in the latest Spring: Spring fundamentals: Spring IoC container, Spring AOP/ AspectJ, and more. Spring enterprise: Spring Java EE integration, Spring Integration, Spring Batch, Spring Remoting, messaging, transactions, and working with big data and the cloud using Hadoop and MongoDB. Spring web: Spring MVC, other dynamic scripting, integration with the popular Grails Framework (and Groovy), REST/web services, and more This book guides you step-by-step through topics using complete and real-world code examples. When you start a new project, you can consider copying the code and configuration files from this book, and then modifying them for your needs. This can save you a great deal of work over creating a project from scratch!

MapReduce Design Patterns Donald Miner 2012-11-21 Until now, design patterns for the MapReduce framework have been scattered among various research papers, blogs, and books. This handy guide brings together a unique collection of valuable MapReduce patterns that will save you time and effort regardless of the domain, language, or development framework you're using. Each pattern is explained in context, with pitfalls and caveats clearly identified to help you avoid common design mistakes when modeling your big data architecture. This book also provides a complete overview of MapReduce that explains its origins and implementations, and why design patterns are so important. All code examples are written for Hadoop. Summarization patterns: get a top-level view by summarizing and grouping data Filtering patterns: view data subsets such as records generated from one user Data organization patterns: reorganize data to work with other systems, or to make MapReduce analysis easier Join patterns: analyze different datasets together to discover interesting relationships Metapatterns: piece together several patterns to solve multi-stage problems, or to perform several analytics in the same job Input and output patterns: customize the way you use Hadoop to load or store data "A clear exposition of MapReduce programs for common data processing patterns—this book is indespensible for anyone using Hadoop." --Tom White, author of Hadoop: The Definitive Guide

PySpark Cookbook Denny Lee 2018-06-29 Combine the power of Apache Spark and Python to build effective big data applications Key Features Perform effective data processing, machine learning, and analytics using PySpark Overcome challenges in developing and deploying Spark solutions using Python Explore recipes for efficiently combining Python and Apache Spark to process data Book Description Apache Spark is an open source framework for efficient cluster computing with a strong interface for data parallelism and fault tolerance. The PySpark Cookbook presents effective and time-saving recipes for leveraging the power of Python and putting it to use in the Spark ecosystem. You'll start by learning the Apache Spark architecture and how to set up a Python environment for Spark. You'll then get familiar with the modules available in PySpark and start using them effortlessly. In addition to this, you'll discover how to abstract data with RDDs and DataFrames, and understand the streaming capabilities of PySpark. You'll then move on to using ML and MLlib in order to solve any problems related to the machine learning capabilities of PySpark and use GraphFrames to solve graph-processing problems. Finally, you will explore how to deploy your applications to the cloud using the spark-submit command. By the end of this book, you will be able to use the Python API for Apache Spark to solve any problems associated with building data-intensive applications. What you will learn Configure a local instance of PySpark in a virtual environment Install and configure Jupyter in local and multi-node environments Create DataFrames from JSON and a dictionary using pyspark.sql Explore regression and clustering models available in the ML module Use DataFrames to transform data used for modeling Connect to PubNub and perform aggregations on streams Who this book is for The PySpark Cookbook is for you if you are a Python developer looking for hands-on recipes for using the Apache Spark 2.x ecosystem in the best possible way. A thorough understanding of Python (and some familiarity with Spark) will help you get the best out of the book.

Expert Hadoop 2 Administration Sam R. Alapati 2016-11-29 This is the eBook of the printed book and may not include any media, website access codes, or print supplements that may come packaged with the bound book. The Comprehensive, Up-to-Date Apache Hadoop Administration Handbook and Reference "Sam Alapati has worked with production Hadoop clusters for six years. His unique depth of experience has enabled him to write the go-to resource for all administrators looking to spec, size, expand, and secure production Hadoop clusters of any size." —Paul Dix, Series Editor In Expert Hadoop® Administration, leading Hadoop administrator Sam R. Alapati brings together authoritative knowledge for creating, configuring, securing, managing, and optimizing production Hadoop clusters in any environment. Drawing on his experience with large-scale Hadoop administration, Alapati integrates action-oriented advice with carefully researched explanations of both problems and solutions. He covers an unmatched range of topics and offers an unparalleled collection of realistic examples. Alapati demystifies complex Hadoop environments, helping you understand exactly what happens behind the scenes when you administer your cluster. You'll gain unprecedented insight as you walk through building clusters from scratch and configuring high availability, performance, security, encryption, and other key attributes. The high-value administration skills you learn here will be indispensable no matter what Hadoop distribution you use or what Hadoop applications you run. Understand Hadoop's architecture from an administrator's standpoint Create simple and fully distributed clusters Run MapReduce and Spark applications in a Hadoop cluster Manage and protect Hadoop data and high availability Work with HDFS commands, file permissions, and storage management Move data, and use YARN to allocate resources and schedule jobs Manage job workflows with Oozie and Hue Secure, monitor, log, and optimize Hadoop Benchmark and troubleshoot Hadoop

Hadoop: The Definitive Guide Tom White 2012-05-10 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

Mastering Hadoop 3 Chanchal Singh 2019-02-28 A comprehensive guide to mastering the most advanced Hadoop 3 concepts Key FeaturesGet to grips with the newly introduced features and capabilities of Hadoop 3Crunch and process data using MapReduce, YARN, and a host of tools within the Hadoop ecosystemSharpen your Hadoop skills with real-world case studies and codeBook Description Apache Hadoop is one of the most popular big data solutions for distributed storage and for processing large chunks of data. With Hadoop 3, Apache promises to provide a high-performance, more fault-tolerant, and highly efficient big data processing platform, with a focus on improved scalability and increased efficiency. With this guide, you'll understand advanced concepts of the Hadoop ecosystem tool. You'll learn how Hadoop works internally, study advanced concepts of different ecosystem tools, discover solutions to real-world use cases, and understand how to secure your cluster. It will then walk you through HDFS, YARN, MapReduce, and Hadoop 3 concepts. You'll be able to address common challenges like using Kafka efficiently, designing low latency, reliable message delivery Kafka systems, and handling high data volumes. As you advance, you'll discover how to address major challenges when building an enterprise-grade messaging system, and how to use different stream processing systems along with Kafka to fulfil your enterprise goals. By the end of this book, you'll have a complete understanding of how components in the Hadoop ecosystem are effectively integrated to implement a fast and reliable data pipeline, and you'll be equipped to tackle a range of real-world problems in data pipelines. What you will learnGain an in-depth understanding of distributed computing using Hadoop 3Develop enterprise-grade applications using Apache Spark, Flink, and moreBuild scalable and high-performance Hadoop data pipelines with security, monitoring, and data governanceExplore batch data processing patterns and how to model data in HadoopMaster best practices for enterprises using, or planning to use, Hadoop 3 as a data platformUnderstand security aspects of Hadoop, including authorization and authenticationWho this book is for If you want to become a big data professional by mastering the advanced concepts of Hadoop, this book is for you. You'll also find this book useful if you're a Hadoop professional looking to strengthen your knowledge of the Hadoop ecosystem. Fundamental knowledge of the Java programming language and basics of Hadoop is necessary to get started with this book.

Hadoop 2 Quick-Start Guide Douglas Eadline 2015-10-28 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

Go Web Development Cookbook Arpit Aggarwal 2018-04-23 86 recipes on how to build fast, scalable, and powerful web services and applications with Go Key Features Become proficient in RESTful web services Build scalable, high-performant web applications in Go Get acquainted with Go frameworks for web development Book Description Go is an open source programming language that is designed to scale and support concurrency at the language level. This gives you the liberty to write large concurrent web applications with ease. From creating web application to deploying them on Amazon Cloud Services, this book will be your one-stop guide to learn web development in Go. The Go Web Development Cookbook teaches you how to create REST services, write microservices, and deploy Go Docker containers. Whether you are new to programming or a professional developer, this book will help get you up to speed with web development in Go. We will focus on writing modular code in Go; in-depth informative examples build the base, one step at a time. You will learn how to create a server, work with static files, SQL, NoSQL databases, and Beego. You will also learn how to create and secure REST services, and create and deploy Go web application and Go Docker containers on Amazon Cloud Services. By the end of the book, you will be able to apply the skills you've gained in Go to create and explore web applications in any domain. What you will learn Create a simple HTTP and TCP web server and understand how it works Explore record in a MySQL and MongoDB database Write and consume RESTful web service in Go Invent microservices in Go using Micro – a microservice toolkit Create and Deploy the Beego application with Nginx Deploy Go web application and Docker containers on an AWS EC2 instance Who this book is for This book is for Go developers interested in learning how to use Go to build powerful web applications. A background in web development is expected.

Practical Data Analysis Hector Cuesta 2013-10-22 Each chapter of the book quickly introduces a key 'theme' of Data Analysis, before immersing you in the practical aspects of each theme. You'll learn quickly how to perform all aspects of Data Analysis.Practical Data Analysis is a book ideal for home and small business users who want to slice & dice the data they have on hand with minimum hassle.

Data Algorithms Mahmoud Parsian 2015-07-13 If you are ready to dive into the MapReduce framework for processing large datasets, this practical book takes you step by step through the algorithms and tools you need to build distributed MapReduce applications with Apache Hadoop or Apache Spark. Each chapter provides a recipe for solving a massive computational problem, such as building a recommendation system. You'll learn how to implement the appropriate MapReduce solution with code that you can use in your projects. Dr. Mahmoud Parsian covers basic design patterns, optimization techniques, and data mining and machine learning solutions for problems in bioinformatics, genomics, statistics, and social network analysis. This book also includes an overview of MapReduce, Hadoop, and Spark. Topics include: Market basket analysis for a large set of transactions Data mining algorithms (K-means, KNN, and Naive Bayes) Using huge genomic data to sequence DNA and RNA Naive Bayes theorem and Markov chains for data and market prediction Recommendation algorithms and pairwise document similarity Linear regression, Cox regression, and Pearson correlation Allelic frequency and mining DNA Social network analysis (recommendation systems, counting triangles, sentiment analysis)

Raspberry Pi □□□□□□□(□□□) Richard Grimmett 2014-09-19 □□□□□□□□□□□□□□□□□□ □□Raspberry Pi□□□ □□□□□□□□□□□□□Raspberry Pi□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□GPS□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□ □□□□□□□□Ubuntu Linux□□□□□Raspberry Pi □□□□□□□□□Open CV□□□□□□□□□□□ □□□□□□□□□□□□□□□□□□□□ □□□□□□□□□GPS□□□□ □□□□□USB□□□□□□□□□□ □□□□□□□□□□□□□□□□□□ □□□□□□□□□□□□□□□□□ □□□□□□□□□□□□□□ #□□□□ GOTOP Information Inc.

Hbase Administration Cookbook Yifeng Jiang 2012-08-16 As part of Packt's cookbook series, each recipe offers a practical, step-by-step solution to common problems found in HBase administration. This book is for HBase administrators, developers, and will even help Hadoop administrators. You are not required to have HBase experience, but are expected to have a basic understanding of Hadoop and MapReduce.

Apache Spark 2.x Cookbook Rishi Yadav 2017-05-31 Over 70 recipes to help you use Apache Spark as your single big data computing platform and master its libraries About This Book This book contains recipes on how to use Apache Spark as a unified compute engine Cover how to connect various source systems to Apache Spark Covers various parts of machine learning including supervised/unsupervised learning & recommendation engines Who This Book Is For This book is for data engineers, data scientists, and those who want to implement Spark for real-time data processing. Anyone who is using Spark (or is planning to) will benefit from this book. The book assumes you have a basic knowledge of Scala as a programming language. What You Will Learn Install and configure Apache Spark with various cluster managers & on AWS Set up a development environment for Apache Spark including Databricks Cloud notebook Find out how to operate on data in Spark with schemas Get to grips with real-time streaming analytics using Spark Streaming & Structured Streaming Master supervised learning and unsupervised learning using MLlib Build a recommendation engine using MLlib Graph processing using GraphX and GraphFrames libraries Develop a set of common applications or project types, and solutions that solve complex big data problems In Detail While Apache Spark 1.x gained a lot of traction and adoption in the early years, Spark 2.x delivers notable improvements in the areas of API, schema awareness, Performance, Structured Streaming, and simplifying building blocks to build better, faster, smarter, and more accessible big data applications. This book uncovers all these features in the form of structured recipes to analyze and mature large and complex sets of data. Starting with installing and configuring Apache Spark with various cluster managers, you will learn to set up development environments. Further on, you will be introduced to working with RDDs, DataFrames and Datasets to operate on schema aware data, and real-time streaming with various sources such as Twitter Stream and Apache Kafka. You will also work through recipes on machine learning, including supervised learning, unsupervised learning & recommendation engines in Spark. Last but not least, the final few chapters delve deeper into the concepts of graph processing using GraphX, securing your implementations, cluster optimization, and troubleshooting. Style and approach This book is packed with intuitive recipes supported with line-by-line explanations to help you understand Spark 2.x's real-time processing capabilities and deploy scalable big data solutions. This is a valuable resource for data scientists and those working on large-scale data projects.

ETL with Azure Cookbook Christian Coté 2020-09-30 Explore the latest Azure ETL techniques both on-premises and in the cloud using Azure services such as SQL Server Integration Services (SSIS), Azure Data Factory, and Azure Databricks Key FeaturesUnderstand the key components of an ETL solution using Azure Integration ServicesDiscover the common and not-so-common challenges faced while creating modern and scalable ETL solutionsProgram and extend your packages to develop efficient data integration and data transformation solutionsBook Description ETL is one of the most common and tedious procedures for moving and processing data from one database to another. With the help of this book, you will be able to speed up the process by designing effective ETL solutions using the Azure services available for handling and transforming any data to suit your requirements. With this cookbook, you'll become well versed in all the features of SQL Server Integration Services (SSIS) to perform data migration and ETL tasks that integrate with Azure. You'll learn how to transform data in Azure and understand how legacy systems perform ETL on-premises using SSIS. Later chapters will get you up to speed with connecting and retrieving data from SQL Server 2019 Big Data Clusters, and even show you how to extend and customize the SSIS toolbox using custom-developed tasks and transforms. This ETL book also contains practical recipes for moving and transforming data with Azure services, such as Data Factory and Azure Databricks, and lets you explore various options for migrating SSIS packages to Azure. Toward the end, you'll find out how to profile data in the cloud and automate service creation with Business Intelligence Markup Language (BIML). By the end of this book, you'll have developed the skills you need to create and automate ETL solutions on-premises as well as in Azure. What you will learnExplore ETL and how it is different from ELTMove and transform various data sources with Azure ETL and ELT servicesUse SSIS 2019 with Azure HDInsight clustersDiscover how to query SQL Server 2019 Big Data Clusters hosted in AzureMigrate SSIS solutions to Azure and solve key challenges associated with itUnderstand why data profiling is crucial and how to implement it in Azure DatabricksGet to grips with BIML and learn how it applies to SSIS and Azure Data Factory solutionsWho this book is for This book is for data warehouse architects, ETL developers, or anyone who wants to build scalable ETL applications in Azure. Those looking to extend their existing on-premise ETL applications to use big data and a variety of Azure services or others interested in migrating existing on-premise solutions to the Azure cloud platform will also find the book useful. Familiarity with SQL Server services is necessary to get the most out of this book.

HBase High Performance Cookbook Ruchir Choudhry 2017-01-31 Exciting projects that will teach you how complex data can be exploited to gain maximum insights About This Book Architect a good HBase cluster for a very large distributed system Get to grips with the concepts of performance tuning with HBase A practical guide full of engaging recipes and attractive screenshots to enhance your system's performance Who This Book Is For This book is intended for developers and architects who want to know all about HBase at a hands-on level. 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. What You Will Learn Configure HBase from a high performance perspective Grab data from various RDBMS/Flat files into the HBASE systems Understand table design and perform CRUD operations Find out how the communication between the client and server happens in HBase Grasp when to use and avoid MapReduce and how to perform various tasks with it Get to know the concepts of scaling with HBase through practical examples Set up Hbase in the Cloud for a small scale environment Integrate HBase with other tools including ElasticSearch In Detail Apache HBase is a non-relational NoSQL database management system that runs on top of HDFS. It is an open source, disturbed, versioned, column-oriented store and is written in Java to provide random real-time access to big Data. We'll start off by ensuring you have a solid understanding the basics of HBase, followed by giving you a thorough explanation of architecting a HBase cluster as per our project specifications. Next, we will explore the scalable structure of tables and we will be able to communicate with the HBase client. After this, we'll show you the intricacies of MapReduce and the art of performance tuning with HBase. Following this, we'll explain the concepts pertaining to scaling with HBase. Finally, you will get an understanding of how to integrate HBase with other tools such as ElasticSearch. By the end of this book, you will have learned enough to exploit HBase for boost system performance. Style and approach This book is intended for software quality assurance/testing professionals, software project managers, or software developers with prior experience in using Selenium and Java to test web-based applications. This books also provides examples for C#, Python, and Ruby users.

Hadoop 2.x Administration Cookbook Gurmukh Singh 2017-05-26 Over 100 practical recipes to help you become an expert Hadoop administrator About This Book Become an expert Hadoop administrator and perform tasks to optimize your Hadoop Cluster Import and export data into Hive and use Oozie to manage workflow. Practical recipes will help you plan and secure your Hadoop cluster, and make it highly available Who This Book Is For If you are a system administrator with a basic understanding of Hadoop and you want to get into Hadoop administration, this book is for you. It's also ideal if you are a Hadoop administrator who wants a quick reference guide to all the Hadoop administration-related tasks and solutions to commonly occurring problems What You Will Learn Set up the Hadoop architecture to run a Hadoop cluster smoothly Maintain a Hadoop cluster on HDFS, YARN, and MapReduce Understand high availability with Zookeeper and Journal Node Configure Flume for data ingestion and Oozie to run various workflows Tune the Hadoop cluster for optimal performance Schedule jobs on a Hadoop cluster using the Fair and Capacity scheduler Secure your cluster and troubleshoot it for various common pain points In Detail Hadoop enables the distributed storage and processing of large datasets across clusters of computers. Learning how to administer Hadoop is crucial to exploit its unique features. With this book, you will be able to overcome common problems encountered in Hadoop administration. The book begins with laying the foundation by showing you the steps needed to set up a Hadoop cluster and its various nodes. You will get a better understanding of how to maintain Hadoop cluster, especially on the HDFS layer and using YARN and MapReduce. Further on, you will explore durability and high availability of a Hadoop cluster. You'll get a better understanding of the schedulers in Hadoop and how to configure and use them for your tasks. You will also get hands-on experience with the backup and recovery options and the performance tuning aspects of Hadoop. Finally, you will get a better understanding of troubleshooting, diagnostics, and best practices in Hadoop administration. By the end of this book, you will have a proper understanding of working with Hadoop clusters and will also be able to secure, encrypt it, and configure auditing for your Hadoop clusters. Style and approach This book contains short recipes that will help you run a Hadoop cluster efficiently. The recipes are solutions to real-life problems that administrators encounter while working with a Hadoop cluster

Apache Mesos Cookbook David Blomquist 2017-08-02 Over 50 recipes on the core features of Apache Mesos and running big data frameworks in Mesos About This Book Learn to install and configure Mesos to suit the needs of your organization Follow step-by-step instructions to deploy application frameworks on top of Mesos, saving you many hours of research and trial and error Use this practical guide packed with powerful recipes to implement Mesos and easily integrate it with other application frameworks Who This Book Is For This book is for system administrators, engineers, and big data programmers. Basic experience with big data technologies such as Hadoop or Spark would be useful but is not essential. A working knowledge of Apache Mesos is expected. What You Will Learn Set up Mesos on different operating systems Use the Marathon and Chronos frameworks to manage multiple applications Work with Mesos and Docker Integrate Mesos with Spark and other big data frameworks Use networking features in Mesos for effective communication between containers Configure Mesos for high availability using Zookeeper Secure your Mesos clusters with SASL and Authorization ACLs Solve everyday problems and discover the best practices In Detail Apache Mesos is open source cluster sharing and management software. Deploying and managing scalable applications in large-scale clustered environments can be difficult, but Apache Mesos makes it easier with efficient resource isolation and sharing across application frameworks. The goal of this book is to guide you through the practical implementation of the Mesos core along with a number of Mesos supported frameworks. You will begin by installing Mesos and then learn how to configure clusters and maintain them. You will also see how to deploy a cluster in a production environment with high availability using Zookeeper. Next, you will get to grips with using Mesos, Marathon, and Docker to build and deploy a PaaS. You will see how to schedule jobs with Chronos. We'll demonstrate how to integrate Mesos with big data frameworks such as Spark, Hadoop, and Storm. Practical solutions backed with clear examples will also show you how to deploy elastic big data jobs. You will find out how to deploy a scalable continuous integration and delivery system on Mesos with Jenkins. Finally, you will configure and deploy a highly scalable distributed search engine with ElasticSearch. Throughout the course of this book, you will get to know tips and tricks along with best practices to follow when working with Mesos. Style and approach This step-by-step guide is packed with powerful recipes on using Apache Mesos and shows its integration with containers and big data frameworks.

Cloudera Administration Handbook Rohit Menon 2014-07-18 An easy-to-follow Apache Hadoop administrator's guide filled with practical screenshots and explanations for each step and configuration. This book is great for administrators interested in setting up and managing a large Hadoop cluster. If you are an administrator, or want to be an administrator, and you are ready to build and maintain a production-level cluster running CDH5, then this book is for you.

Kafka: The Definitive Guide Neha Narkhede 2017-08-31 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

Trino: The Definitive Guide Matt Fuller 2021-04-14 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

Apache Sqoop Cookbook Kathleen Ting 2013-07-02 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

Spark Cookbook Rishi Yadav 2015-07-27 By introducing in-memory persistent storage, Apache Spark eliminates the need to store intermediate data in filesystems, thereby increasing processing speed by up to 100 times. This book will focus on how to analyze large and complex sets of data. Starting with installing and configuring Apache Spark with various cluster managers, you will cover setting up development environments. You will then cover various recipes to perform interactive queries using Spark SQL and real-time streaming with various sources such as Twitter Stream and Apache Kafka. You will then focus on machine learning, including supervised learning, unsupervised learning, and recommendation engine algorithms. After mastering graph processing using GraphX, you will cover various recipes for cluster optimization and troubleshooting.

Hadoop Operations Eric Sammer 2012-09-26 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

Pentaho Data Integration Cookbook Alex Meadows 2013-12-02 Pentaho Data Integration Cookbook Second Edition is written in a cookbook format, presenting examples in the style of recipes.This allows you to go directly to your topic of interest, or follow topics throughout a chapter to gain a thorough in-depth knowledge.Pentaho Data Integration Cookbook Second Edition is designed for developers who are familiar with the basics of Kettle but who wish to move up to the next level.It is also aimed at advanced users that want to learn how to use the new features of PDI as well as and best practices for working with Kettle.

HDInsight Essentials - Second Edition Rajesh Nadipalli 2015-01-27 If you want to discover one of the latest tools designed to produce stunning Big Data insights, this book features everything you need to get to grips with your data. Whether you are a data architect, developer, or a business strategist, HDInsight adds value in everything from development, administration, and reporting.

IBM Cognos Dynamic Query Nigel Campbell 2013-09-12 This IBM® Redbooks® publication explains how IBM Cognos® Business Intelligence (BI) administrators, authors, modelers, and power users can use the dynamic query layer effectively. It provides guidance for determining which technology within the dynamic query layer can best satisfy your business requirements. Administrators can learn how to tune the query service effectively and preferred practices for managing their business intelligence content. This book includes information about metadata modeling of relational data sources with IBM Cognos Framework Manager. It includes considerations that can help you author high-performing applications that satisfy analytical requirements of users. This book provides guidance for troubleshooting issues related to the dynamic query layer of Cognos BI. Related documents: Solution Guide : Big Data Analytics with IBM Cognos BI Dynamic Query Blog post : IBM Cognos Dynamic Query Extensibility

Ceph Cookbook Vikhyat Umrao 2017-11-24 Over 100 effective recipes to help you design, implement, and troubleshoot manage the software-defined and massively scalable Ceph storage system. About This Book Implement a Ceph cluster successfully and learn to manage it. Recipe based approach in learning the most efficient software defined storage system Implement best practices on improving efficiency and security of your storage cluster Learn to troubleshoot common issues experienced in a Ceph cluster Who This Book Is For This book is targeted at storage and cloud engineers, system administrators, or anyone who is interested in building software defined storage, to power your cloud or virtual infrastructure. If you have basic knowledge of GNU/Linux and storage systems, with no experience of software defined storage solutions and Ceph, but eager to learn then this book is for you What You Will Learn Understand, install, configure, and manage the Ceph storage system Get to grips with performance tuning and benchmarking, and learn practical tips to help run Ceph in production Integrate Ceph with OpenStack Cinder, Glance, and Nova components Deep dive into Ceph object storage, including S3, Swift, and Keystone integration Configure a disaster recovery solution with a Ceph Multi-Site V2 gateway setup and RADOS Block Device mirroring Gain hands-on experience with Ceph Metrics and VSM for cluster monitoring Familiarize yourself with Ceph operations such as maintenance, monitoring, and troubleshooting Understand advanced topics including erasure-coding, CRUSH map, cache pool, and general Ceph cluster maintenance In Detail Ceph is a unified distributed storage system designed for reliability and scalability. This technology has been transforming the software-defined storage industry and is evolving rapidly as a leader with its wide range of support for popular cloud platforms such as OpenStack, and CloudStack, and also for virtualized platforms. Ceph is backed by Red Hat and has been developed by community of developers which has gained immense traction in recent years. This book will guide you right from the basics of Ceph , such as creating blocks, object storage, and filesystem access, to advanced concepts such as cloud integration solutions. The book will also cover practical and easy to implement recipes on CephFS, RGW, and RBD with respect to the major stable release of Ceph Jewel. Towards the end of the book, recipes based on troubleshooting and best practices will help you get to grips with managing Ceph storage in a production environment. By the end of this book, you will have practical, hands-on experience of using Ceph efficiently for your storage requirements. Style and approach This step-by-step guide is filled with practical tutorials, making complex scenarios easy to understand.

Apache Kafka 1.0 Cookbook Raúl Estrada 2017-12-22 Simplify real-time data processing by leveraging the power of Apache Kafka 1.0 Key Features Use Kafka 1.0 features such as Confluent platforms and Kafka streams to build efficient streaming data applications to handle and process your data Integrate Kafka with other Big Data tools such as Apache Hadoop, Apache Spark, and more Hands-on recipes to help you design, operate, maintain, and secure your Apache Kafka cluster with ease Book Description Apache Kafka provides a unified, high-throughput, low-latency platform to handle real-time data feeds. This book will show you how to use Kafka efficiently, and contains practical solutions to the common problems that developers and administrators usually face while working with it. This practical guide contains easy-to-follow recipes to help you set up, configure, and use Apache Kafka in the best possible manner. You will use Apache Kafka Consumers and Producers to build effective real-time streaming applications. The book covers the recently released Kafka version 1.0, the Confluent Platform and Kafka Streams. The programming aspect covered in the book will teach you how to perform important tasks such as message validation, enrichment and composition.Recipes focusing on optimizing the performance of your Kafka cluster, and integrate Kafka with a variety of third-party tools such as Apache Hadoop, Apache Spark, and Elasticsearch will help ease your day to day collaboration with Kafka greatly. Finally, we cover tasks related to monitoring and securing your Apache Kafka cluster using tools such as Ganglia and Graphite. If you're looking to become the go-to person in your organization when it comes to working with Apache Kafka, this book is the only resource you need to have. What you will learn -Install and configure Apache Kafka 1.0 to get optimal performance -Create and configure Kafka Producers and Consumers -Operate your Kafka clusters efficiently by implementing the mirroring technique -Work with the new Confluent platform and Kafka streams, and achieve high availability with Kafka -Monitor Kafka using tools such as Graphite and Ganglia -Integrate Kafka with third-party tools such as Elasticsearch, Logstash, Apache Hadoop, Apache Spark, and more Who this book is for This book is for developers and Kafka administrators who are looking for quick, practical solutions to problems encountered while operating, managing or monitoring Apache Kafka. If you are a developer, some knowledge of Scala or Java will help, while for administrators, some working knowledge of Kafka will be useful.

Hadoop Real-World Solutions Cookbook Tanmay Deshpande 2016-03-31 Over 90 hands-on recipes to help you learn and master the intricacies of Apache Hadoop 2.X, YARN, Hive, Pig, Oozie, Flume, Sqoop, Apache Spark, and Mahout About This Book Implement outstanding Machine Learning use cases on your own analytics models and processes. Solutions to common problems when working with the Hadoop ecosystem. Step-by-step implementation of end-to-end big data use cases. Who This Book Is For Readers who have a basic knowledge of big data systems and want to advance their knowledge with hands-on recipes. What You Will Learn Installing and maintaining Hadoop 2.X cluster and its ecosystem. Write advanced Map Reduce programs and understand design patterns. Advanced Data Analysis using the Hive, Pig, and Map Reduce programs. Import and export data from various sources using Sqoop and Flume. Data storage in various file formats such as Text, Sequential, Parquet, ORC, and RC Files. Machine learning principles with libraries such as Mahout Batch and Stream data processing using Apache Spark In Detail Big data is the current requirement. Most organizations produce huge amount of data every day. With the arrival of Hadoop-like tools, it has become easier for everyone to solve big data problems with great efficiency and at minimal cost. Grasping Machine Learning techniques will help you greatly in building predictive models and using this data to make the right decisions for your organization. Hadoop Real World Solutions Cookbook gives readers insights into learning and mastering big data via recipes. The book not only clarifies most big data tools in the market but also provides best practices for using them. The book provides recipes that are based on the latest versions of Apache Hadoop 2.X, YARN, Hive, Pig, Sqoop, Flume, Apache Spark, Mahout and many more such ecosystem tools. This real-world-solution cookbook is packed with handy recipes you can apply to your own everyday issues. Each chapter provides in-depth recipes that can be referenced easily. This book provides detailed practices on the latest technologies such as YARN and Apache Spark. Readers will be able to consider themselves as big data experts on completion of this book. This guide is an invaluable tutorial if you are planning to implement a big data warehouse for your business. Style and approach An easy-to-follow guide that walks you through world of big data. Each tool in the Hadoop ecosystem is explained in detail and the recipes are placed in such a manner that readers can implement them sequentially. Plenty of reference links are provided for advanced reading.

Hadoop Operations and Cluster Management Cookbook Shumin Guo 2013 Solve specific problems using individual self-contained code recipes, or work through the book to develop your capabilities. This book is packed with easy-to-follow code and commands used for illustration, which makes your learning curve easy and quick.If you are a Hadoop cluster system administrator with Unix/Linux system management experience and you are looking to get a good grounding in how to set up and manage a Hadoop cluster, then this book is for you. It's assumed that you will have some experience in Unix/Linux command line already, as well as being familiar with network communication basics.

Apache Mesos Cookbook David Blomquist 2017-04-28 Over 50 recipes on the core features of Apache Mesos and running big data frameworks in MesosAbout This Book* Learn to install and configure Mesos to suit the needs of your organization* Follow step-by-step instructions to deploy application frameworks on top of Mesos, saving you many hours of research and trial and error* Use this practical guide packed with powerful recipes to implement Mesos and easily integrate it with other application frameworksWho This Book Is ForThis book is for system administrators, engineers, and big data programmers. Basic experience with big data technologies such as Hadoop or Spark would be useful but is not essential. A working knowledge of Apache Mesos is expected.What you will learn* Set up Mesos on different operating systems* Use the Marathon and Chronos frameworks to manage multiple applications* Work with Mesos and Docker* Integrate Mesos with Spark and other big data frameworks* Use networking features in Mesos for effective communication between containers* Configure Mesos for high availability using Zookeeper* Secure your Mesos clusters with SASL and Authorization ACLs* Solve everyday problems and discover the best practicesIn DetailApache Mesos is open source cluster sharing and management software. Deploying and managing scalable applications in large-scale clustered environments can be difficult, but Apache Mesos makes it easier with efficient resource isolation and sharing across application frameworks.The goal of this book is to guide you through the practical implementation of the Mesos core along with a number of Mesos supported frameworks. You will begin by installing Mesos and then learn how to configure clusters and maintain them. You will also see how to deploy a cluster in a production environment with high availability using Zookeeper.Next, you will get to grips with using Mesos, Marathon, and Docker to build and deploy a PaaS. You will see how to schedule jobs with Chronos. We'll demonstrate how to integrate Mesos with big data frameworks such as Spark, Hadoop, and Storm. Practical solutions backed with clear examples will also show you how to deploy elastic big data jobs.You will find out how to deploy a scalable continuous integration and delivery system on Mesos with Jenkins. Finally, you will configure and deploy a highly scalable distributed search engine with ElasticSearch.Throughout the course of this book, you will get to know tips and tricks along with best practices to follow when working with Mesos.

Big Data Analytics with R and Hadoop Vignesh Prajapati 2013-11-25 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.

Apache Karaf Cookbook Achim Nierbeck 2014-08-25 This book is intended for developers who have some familiarity with Apache Karaf and who want a quick reference for practical, proven tips on how to perform common tasks such as configuring Pax modules deployed in Apache Karaf, Extending HttpService with Apache Karaf. You should have working knowledge of Apache karaf, as the book provides a deeper understanding of the capabilities of Apache Karaf.

Hadoop Beginner's Guide Garry Turkington 2013-02-22 Data is arriving faster than you can process it and the overall volumes keep growing at a rate that keeps you awake at night. Hadoop can help you tame the data beast. Effective use of Hadoop however requires a mixture of programming, design, and system administration skills. "Hadoop Beginner's Guide" removes the mystery from Hadoop, presenting Hadoop and related technologies with a focus on building working systems and getting the job done, using cloud services to do so when it makes sense. From basic concepts and initial setup through developing applications and keeping the system running as the data grows, the book gives the understanding needed to effectively use Hadoop to solve real world problems. Starting with the basics of installing and configuring Hadoop, the book explains how to develop applications, maintain the system, and how to use additional products to integrate with other systems. While learning different ways to develop applications to run on Hadoop the book also covers tools such as Hive, Sqoop, and Flume that show how Hadoop can be integrated with relational databases and log collection. In addition to examples on Hadoop clusters on Ubuntu uses of cloud services such as Amazon, EC2 and Elastic MapReduce are covered.

SQL Server 2017 Integration Services Cookbook Christian Cote 2017-06-30 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.

Kubernetes Cookbook Sébastien Goasguen 2018-02-14 If your organization is preparing to move toward a cloud-native computing architecture, this cookbook shows you how to successfully use Kubernetes, the de-facto standard for automating the deployment, scaling, and management of containerized applications. With more than 80 proven recipes, developers, system administrators, and architects will quickly learn how to get started with Kubernetes and understand its powerful API. Through the course of the book, authors Sébastien Goasguen and Michael Hausenblas provide several detailed solutions for installing, interacting with, and using Kubernetes in development and production. You'll learn how to adapt the system to your particular needs and become familiar with the wider Kubernetes ecosystem. Each standalone chapter features recipes written in O'Reilly's popular problem-solution-discussion format. Recipes in this cookbook focus on: Creating a Kubernetes cluster Using the Kubernetes command-line interface Managing fundamental workload types Working with services Exploring the Kubernetes API Managing stateful and non-cloud native apps Working with volumes and configuration data Cluster-level and application-level scaling Securing your applications Monitoring and logging Maintenance and troubleshooting.

Programming 101 Jeanine Meyer 2018-06-15 Understand the importance of programming, even if you’ve never programmed before! This book will teach you the basics of programming using the Processing programming language. You will create your own Processing sketches, using personal images, themes, or hobbies that you enjoy. The chapters in the book will demonstrate the process of programming, starting with formulating an idea, planning, building on past projects, and refining the work, similar to writing an essay or composing a song. This approach will guide you to make use of logic and mathematics to produce beautiful effects. The term for program in Processing is sketch, though the sketches featured in this book are far more than static drawings; they incorporate interaction, animation, video, audio, and accessing files on the local computer and on the Web. Technical features are introduced and explained in the context of complete examples: games (Snake, Hangman, jigsaw, slingshot), making a collage of family images and video clips, preparing directions for folding an origami model, rotating objects in 3D, and others. Programming is a fun, creative, expressive pursuit. It requires attention to details and can be frustrating, but there is very little that compares to the satisfaction of building a program out of nothing and making it work (or taking an existing program and fixing a problem, or adding a feature and making it better). Programming 101 is your gateway to making this happen. What You Will Learn Gain basic programming skills Build fun and creative programs Use files for making a holiday card Combine videos, images, and graphics in a Processing sketch Who This Book Is For Anyone who has been thinking about trying programming, or has tried, but needs more motivation; anyone who wants to learn about the Processing language.

Azure Databricks Cookbook Phani Raj 2021-09-17 Get to grips with building and productionizing end-to-end big data solutions in Azure and learn best practices for working with large datasets Key FeaturesIntegrate with Azure Synapse Analytics, Cosmos DB, and Azure HDInsight Kafka Cluster to scale and analyze your projects and build pipelinesUse Databricks SQL to run ad hoc queries on your data lake and create dashboardsProductionize a solution using CI/CD for deploying notebooks and Azure Databricks Service to various environmentsBook Description Azure Databricks is a unified collaborative platform for performing scalable analytics in an interactive environment. The Azure Databricks Cookbook provides recipes to get hands-on with the analytics process, including ingesting data from various batch and streaming sources and building a modern data warehouse. The book starts by teaching you how to create an Azure Databricks instance within the Azure portal, Azure CLI, and ARM templates. You'll work through clusters in Databricks and explore recipes for ingesting data from sources, including files, databases, and streaming sources such as Apache Kafka and EventHub. The book will help you explore all the features supported by Azure Databricks for building powerful end-to-end data pipelines. You'll also find out how to build a modern data warehouse by using Delta tables and Azure Synapse Analytics. Later, you'll learn how to write ad hoc queries and extract meaningful insights from the data lake by creating visualizations and dashboards with Databricks SQL. Finally, you'll deploy and productionize a data pipeline as well as deploy notebooks and Azure Databricks service using continuous integration and continuous delivery (CI/CD). By the end of this Azure book, you'll be able to use Azure Databricks to streamline different processes involved in building data-driven apps. What you will learnRead and write data from and to various Azure resources and file formatsBuild a modern data warehouse with Delta Tables and Azure Synapse AnalyticsExplore jobs, stages, and tasks and see how Spark lazy evaluation worksHandle concurrent transactions and learn performance optimization in Delta tablesLearn Databricks SQL and create real-time dashboards in Databricks SQLIntegrate Azure DevOps for version control, deploying, and productionizing solutions with CI/CD pipelinesDiscover how to use RBAC and ACLs to restrict data accessBuild end-to-end data processing pipeline for near real-time data analyticsWho this book is for This recipe-based book is for data scientists, data engineers, big data professionals, and machine learning engineers who want to perform data analytics on their applications. Prior experience of working with Apache Spark and Azure is necessary to get the most out of this book.

PostgreSQL 12 High Availability Cookbook Shaun Thomas 2020-02-25 A comprehensive guide to understanding key techniques for architecture and hardware planning, monitoring, replication, backups, and decoupling Key FeaturesNewly updated edition, covering the latest PostgreSQL 12 features with hands-on industry-driven recipesCreate a PostgreSQL cluster that stays online even when disaster strikesLearn how to avoid costly downtime and data loss that can ruin your businessBook Description Databases are nothing without the data they store. In the event of an outage or technical catastrophe, immediate recovery is essential. This updated edition ensures that you will learn the important concepts related to node architecture design, as well as techniques such as using repmgr for failover automation. From cluster layout and hardware selection to software stacks and horizontal scalability, this PostgreSQL cookbook will help you build a PostgreSQL cluster that will survive crashes, resist data corruption, and grow smoothly with customer demand. You’ll start by understanding how to plan a PostgreSQL database architecture that is resistant to outages and scalable, as it is the scaffolding on which everything rests. With the bedrock established, you'll cover the topics that PostgreSQL database administrators need to know to manage a highly available cluster. This includes configuration, troubleshooting, monitoring and alerting, backups through proxies, failover automation, and other considerations that are essential for a healthy PostgreSQL cluster. Later, you'll learn to use multi-master replication to maximize server availability. Later chapters will guide you through managing major version upgrades without downtime. By the end of this book, you'll have learned how to build an efficient and adaptive PostgreSQL 12 database cluster. What you will learnUnderstand how to protect data with PostgreSQL replication toolsFocus on hardware planning to ensure that your database runs efficientlyReduce database resource contention with connection poolingMonitor and visualize cluster activity with Nagios and the TIG (Telegraf, InfluxDB, Grafana) stack Construct a robust software stack that can detect and avert outagesUse multi-master to achieve an enduring PostgreSQL clusterWho this book is for This book is for Postgres administrators and developers who are looking to build and maintain a highly reliable PostgreSQL cluster. Although knowledge of the new features of PostgreSQL 12 is not required, a basic understanding of PostgreSQL administration is expected.

Spark: The Definitive Guide Bill Chambers 2018-02-08 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

Downloaded from beenews.com on February 5, 2023 by guest