Hadoop Operations And Cluster Management Cookbook

Raspberry Pi 機器人自造專案(電子書) Richard Grimmett 2014-09-19 從零開始，花小錢打造屬於您自己的機器人 機器人的發展神速，現在，只要您有興趣，就能輕鬆取得製作機器人所需的硬體元件以及相關軟體，打造屬於自己的機器人。本書會從啟用Raspberry Pi這塊板子開始。它可以解譯您的指令並讓機器人執行各種動作。跟著本書的每個單元進行實作，假以時日，您就能擁有一台可說、可聽並可用多種令人讚嘆不已的方式來移動的機器人了。跟著本書的步驟導引，您就能應用Raspberry Pi實作出各項有趣的功能，包括語音辨識、仿人語音模擬、圖像辨識、馬達控制、GPS定位以及無線控制。您還可以學到如何整合這些功能來製作專屬的機器人專案。．開箱、供電並用Ubuntu Linux來設定您的Raspberry Pi．在您的專題中應用Open CV來辨識顏色、樣式與動作．加入並控制語音軟體讓您的機器人可以說話．為您的機器人加上GPS定位功能．使用無線USB鍵盤來遙控您的機器人．透過各種外部硬體讓您的機器人動起來．製作可移動、下水甚至飛行的機器人．與您的機器人擺脫纜線的束縛 #碁峰資訊 GOTOP Information Inc.

Datenintensive Anwendungen designen Martin Kleppmann 2018-11-26 Daten stehen heute im Mittelpunkt vieler Herausforderungen im Systemdesign. Dabei sind komplexe Fragen wie Skalierbarkeit, Konsistenz, Zuverlässigkeit, Effizienz und Wartbarkeit zu klären. Darüber hinaus verfügen wir über eine überwältigende Vielfalt an Tools, einschließlich relationaler Datenbanken, NoSQL-Datenspeicher, Stream-und Batchprocessing und Message Broker. Aber was verbirgt sich hinter diesen Schlagworten? Und was ist die richtige Wahl für Ihre Anwendung? In diesem praktischen und umfassenden Leitfaden unterstützt Sie der Autor Martin Kleppmann bei der Navigation durch dieses schwierige Terrain, indem er die Vor-und Nachteile verschiedener Technologien zur Verarbeitung und Speicherung von Daten aufzeigt. Software verändert sich ständig, die Grundprinzipien bleiben aber gleich. Mit diesem Buch lernen Softwareentwickler und -architekten, wie sie die Konzepte in der Praxis umsetzen und wie sie Daten in modernen Anwendungen optimal nutzen können. Inspizieren Sie die Systeme, die Sie bereits verwenden, und erfahren Sie, wie Sie sie effektiver nutzen können Treffen Sie fundierte Entscheidungen, indem Sie die Stärken und Schwächen verschiedener Tools kennenlernen Steuern Sie die notwenigen Kompromisse in Bezug auf Konsistenz, Skalierbarkeit, Fehlertoleranz und Komplexität Machen Sie sich vertraut mit dem Stand der Forschung zu verteilten Systemen, auf denen moderne Datenbanken aufbauen Werfen Sie einen Blick hinter die Kulissen der wichtigsten Onlinedienste und lernen Sie von deren Architekturen

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

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.

Mastering Python Scientific Computing Hemant Kumar Mehta 2015-09-23 A complete guide for Python programmers to master scientific computing using Python APIs and tools About This Book The basics of scientific computing to advanced concepts involving parallel and large scale computation are all covered. Most of the Python APIs and tools used in scientific computing are discussed in detail The concepts are discussed with suitable example programs Who This Book Is For If you are a Python programmer and want to get your hands on scientific computing, this book is for you. The book expects you to have had exposure to various concepts of Python programming. What You Will Learn Fundamentals and components of scientific computing Scientific computing data management Performing numerical computing using NumPy and SciPy Concepts and programming for symbolic computing using SymPy Using the plotting library matplotlib for data visualization Data analysis and visualization using Pandas, matplotlib, and IPython Performing parallel and high performance computing Real-life case studies and best practices of scientific computing In Detail In today's world, along with theoretical and experimental work, scientific computing has become an important part of scientific disciplines. Numerical calculations, simulations and computer modeling in this day and age form the vast majority of both experimental and theoretical papers. In the scientific method, replication and reproducibility are two important contributing factors. A complete and concrete scientific result should be reproducible and replicable. Python is suitable for scientific computing. A large community of users, plenty of help and documentation, a large collection of scientific libraries and environments, great performance, and good support makes Python a great choice for scientific computing. At present Python is among the top choices for developing scientific workflow and the book targets existing Python developers to master this domain using Python. The main things to learn in the book are the concept of scientific workflow, managing scientific workflow data and performing computation on this data using Python. The book discusses NumPy, SciPy, SymPy, matplotlib, Pandas and IPython with several example programs. Style and approach This book follows a hands-on approach to explain the complex concepts related to scientific computing. It details various APIs using appropriate examples.

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.

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.

Architecting Modern Data Platforms Jan Kunigk 2018-12-05 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

Microservices Deployment Cookbook Vikram Murugesan 2017-01-31 Master over 60 recipes to help you deliver complete, scalable, microservice-based solutions and see the improved business results immediately About This Book Adopt microservices-based architecture and deploy it at scale Build your complete microservice architecture using different recipes for different solutions Identify specific tools for specific scenarios and deliver immediate business results, correlate use cases, and adopt them in your team and organization Who This Book Is For This book is for developers, ops, and DevOps professionals who would like to put microservices to work and improve products, services, and operations. Those looking to build and deploy microservices will find this book useful, as well as managers and people at CXO

level looking to adopt microservices in their organization. Prior knowledge of Java is expected. No prior knowledge of microservices is assumed. What You Will Learn Build microservices using Spring Boot, Wildfly Swarm, Dropwizard, and SparkJava Containerize your microservice using Docker Deploy microservices using Mesos/Marathon and Kubernetes Implement service discovery and load balancing using Zookeeper, Consul, and Nginx Monitor microservices using Graphite and Grafana Write stream programs with Kafka Streams and Spark Aggregate and manage logs using Kafka Get introduced to DC/OS, Docker Swarm, and YARN In Detail This book will help any team or organization understand, deploy, and manage microservices at scale. It is driven by a sample application, helping you gradually build a complete microservice-based ecosystem. Rather than just focusing on writing a microservice, this book addresses various other microservice-related solutions: deployments, clustering, load balancing, logging, streaming, and monitoring. The initial chapters offer insights into how web and enterprise apps can be migrated to scalable microservices. Moving on, you'll see how to Dockerize your application so that it is ready to be shipped and deployed. We will look at how to deploy microservices on Mesos and Marathon and will also deploy microservices on Kubernetes. Next, you will implement service discovery and load balancing for your microservices. We'll also show you how to build asynchronous streaming systems using Kafka Streams and Apache Spark. Finally, we wind up by aggregating your logs in Kafka, creating your own metrics, and monitoring the metrics for the microservice. Style and approach This book follows a recipe-driven approach and shows you how to plug and play with all the various pieces, putting them together to build a complete scalable microservice ecosystem. You do not need to study the chapters in order, as you can directly refer to the content you need for your situation.

Data Management in Machine Learning Systems Matthias Boehm 2019-02-25 Large-scale data analytics using machine learning (ML) underpins many modern data-driven applications. ML systems provide means of specifying and executing these ML workloads in an efficient and scalable manner. Data management is at the heart of many ML systems due to data-driven application characteristics, data-centric workload characteristics, and system architectures inspired by classical data management techniques. In this book, we follow this data-centric view of ML systems and aim to provide a comprehensive overview of data management in ML systems for the end-to-end data science or ML lifecycle. We review multiple interconnected lines of work: (1) ML support in database (DB) systems, (2) DB-inspired ML systems, and (3) ML lifecycle systems. Covered topics include: in-database analytics via query generation and user-defined functions, factorized and statistical-relational learning; optimizing compilers for ML workloads; execution strategies and hardware accelerators; data access methods such as compression, partitioning and indexing; resource elasticity and cloud markets; as well as systems for data preparation for ML, model selection, model management, model debugging, and model serving. Given the rapidly evolving field, we strive for a balance between an up-to-date survey of ML systems, an overview of the underlying concepts and techniques, as well as pointers to open research questions. Hence, this book might serve as a starting point for both systems researchers and developers.

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.

Machine Learning Kochbuch Chris Albon 2019-03-22 Python-Programmierer finden in diesem Kochbuch nahezu 200 wertvolle und jeweils in sich abgeschlossene Anleitungen zu Aufgabenstellungen aus dem Bereich des Machine Learning, wie sie für die tägliche Arbeit typisch sind – von der Vorverarbeitung der Daten bis zum Deep Learning. Entwickler, die mit Python und seinen Bibliotheken einschließlich Pandas und Scikit-Learn vertraut sind, werden spezifische Probleme erfolgreich bewältigen – wie etwa Daten laden, Text und numerische Daten behandeln, Modelle auswählen, Dimensionalität reduzieren und vieles mehr. Jedes Rezept enthält Code, den Sie kopieren, zum Testen in eine kleine Beispieldatenmenge einfügen und dann anpassen können, um Ihre eigenen Anwendungen zu konstruieren. Darüber hinaus werden alle Lösungen diskutiert und wichtige Zusammenhänge hergestellt. Dieses Kochbuch unterstützt Sie dabei, den Schritt von der Theorie und den Konzepten hinein in die Praxis zu machen. Es liefert das praktische Rüstzeug, das Sie benötigen, um funktionierende Machine-Learning-Anwendungen zu entwickeln. In diesem Kochbuch finden Sie Rezepte für: Vektoren, Matrizen und Arrays den Umgang mit numerischen und kategorischen Daten, Texten, Bildern sowie Datum und Uhrzeit das Reduzieren der Dimensionalität durch Merkmalsextraktion oder Merkmalsauswahl Modellbewertung und -auswahl lineare und logistische Regression, Bäume und Wälder und k-nächste Nachbarn Support Vector Machine (SVM), naive Bayes, Clustering und neuronale Netze das Speichern und Laden von trainierten Modellen

PySpark Recipes Raju Kumar Mishra 2017-12-09 Quickly find solutions to common programming problems encountered while processing big data. Content is presented in the popular problem-solution format. Look up the programming problem that you want to solve. Read the solution. Apply the solution directly in your own code. Problem solved! PySpark Recipes covers Hadoop and its shortcomings. The architecture of Spark, PySpark, and RDD are presented. You will learn to apply RDD to solve day-to-day big data problems. Python and NumPy are included and make it easy for new learners of PySpark to understand and adopt the model. What You Will Learn Understand the advanced features of PySpark2 and SparkSQL Optimize your code Program SparkSQL with Python Use Spark Streaming and Spark MLlib with Python Perform graph analysis with GraphFrames Who This Book Is For Data analysts, Python programmers, big data enthusiasts

Professional Hadoop Benoy Antony 2016-05-23 The professional's one-stop guide to this open-source, Java-based big data framework Professional Hadoop is the complete reference and resource for experienced developers looking to employ Apache Hadoop in real-world settings. Written by an expert team of certified Hadoop developers, committers, and Summit speakers, this book details every key aspect of Hadoop technology to enable optimal processing of large data sets. Designed expressly for the professional developer, this book skips over the basics of database development to get you acquainted with the framework's processes and capabilities right away. The discussion covers each key Hadoop component individually, culminating in a sample application that brings all of the pieces together to illustrate the cooperation and interplay that make Hadoop a major big data solution. Coverage includes everything from storage and security to computing and user experience, with expert guidance on integrating other software and more. Hadoop is quickly reaching significant market usage, and more and more developers are being called upon to develop big data solutions using the Hadoop framework. This book covers the process from beginning to end, providing a crash course for professionals needing to learn and apply Hadoop quickly. Configure storage, UE, and in-memory computing Integrate Hadoop with other programs including Kafka and Storm Master the fundamentals of Apache Big Top and Ignite Build robust data security with expert tips and advice Hadoop's popularity is largely due to its accessibility. Open-source and written in Java, the framework offers almost no barrier to entry for experienced database developers already familiar with the skills and requirements real-world programming entails. Professional Hadoop gives you the practical information and framework-specific skills you need quickly.

Learning Hunk Dmitry Anoshin 2015-12-31 Visualize and analyze your Hadoop data using Hunk About This Book Explore your data in Hadoop and NoSQL data stores Create and optimize your reporting experience with advanced data visualizations and data analytics A comprehensive developer's guide that helps you create outstanding analytical solutions efficiently Who This Book Is For If you are Hadoop developers who want to build efficient real-time Operation Intelligence Solutions based on Hadoop deployments or various NoSQL data stores using Hunk, this book is for you. Some familiarity with Splunk is assumed. What You Will Learn Deploy and configure Hunk on top of Cloudera Hadoop Create and configure Virtual Indexes for datasets Make your data presentable using the wide variety of data visualization components and knowledge objects Design a data model using Hunk best practices Add more flexibility to your analytics solution via extended SDK and custom visualizations Discover data using MongoDB as a data source Integrate Hunk with AWS Elastic MapReduce to improve scalability In Detail Hunk is the big data analytics platform that lets you rapidly explore, analyse, and visualize data in Hadoop and NoSQL data stores. It provides a single, fluid user experience, designed to show you insights from your big data without the need for specialized skills, fixed schemas, or months of development. Hunk goes beyond typical data analysis methods and gives you the power to rapidly detect patterns and find anomalies across petabytes of raw data. This book focuses on exploring, analysing, and visualizing big data in Hadoop and NoSQL data stores with this powerful full-featured big data analytics platform. You will begin by learning the Hunk architecture and Hunk Virtual Index before moving on to how to easily analyze and visualize data using Splunk Search Language (SPL). Next you will meet Hunk Apps which can easy integrate with NoSQL data stores such as MongoDB or Sqqrl. You will also discover Hunk knowledge objects, build a semantic layer on top of Hadoop, and explore data using the friendly user-interface of Hunk Pivot. You will connect MongoDB and explore data in the data store. Finally, you will go through report acceleration techniques and analyze data in the AWS Cloud. Style and approach A step-by-step guide starting right from the basics and deep diving into the more advanced and technical aspects of Hunk.

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.

Big Data Management Peter Ghavami 2020-11-09 Data analytics is core to business and decision making. The rapid increase in data volume, velocity and variety offers both opportunities and challenges. While open source solutions to store big data, like Hadoop, offer platforms for exploring value and insight from big data, they were not originally developed with data security and governance in mind. Big Data Management discusses numerous policies, strategies and recipes for managing big data. It addresses data security, privacy, controls and life cycle management offering modern principles and open source architectures for successful governance of big data. The author has collected best practices from the world's leading organizations that have successfully implemented big data platforms. The topics discussed cover the entire data management life cycle, data quality, data stewardship, regulatory considerations, data council, architectural and operational models are presented for successful management of big data. The book is a must-read for data scientists, data engineers and corporate leaders who are implementing big data platforms in their organizations.

Scala: Guide for Data Science Professionals Pascal Bugnion 2017-02-24 Scala will be a valuable tool to have on hand during your data science journey for everything from data cleaning to cutting-edge machine learning About This Book Build data science and data engineering solutions with ease An in-depth look at each stage of the data analysis process — from reading and collecting data to distributed analytics Explore a broad variety of data processing, machine learning, and genetic algorithms through diagrams, mathematical formulations, and source code Who This Book Is For This learning path is perfect for those who are comfortable with Scala programming and now want to enter the field of data science. Some knowledge of statistics is expected. What You Will Learn Transfer and filter tabular data to extract features for machine learning Read, clean, transform, and write data to both SQL and NoSQL databases Create Scala web applications that couple with JavaScript libraries such as D3 to create compelling interactive visualizations Load data from HDFS and HIVE with ease Run streaming and graph analytics in Spark for exploratory analysis Bundle and scale up Spark jobs by deploying them into a variety of cluster managers Build dynamic workflows for scientific computing Leverage open source libraries to extract patterns from time series Master probabilistic models for sequential data In Detail Scala is especially good for analyzing large sets of data as the scale of the task doesn't have any significant impact on performance. Scala's powerful functional libraries can interact with databases and build scalable frameworks — resulting in the creation of robust data pipelines. The first module introduces you to Scala libraries to ingest, store, manipulate, process, and visualize data. Using real world examples, you will learn how to design scalable architecture to process and model data — starting from simple concurrency constructs and progressing to actor systems and Apache Spark. After this, you will also learn how to build interactive visualizations with web frameworks. Once you have become familiar with all the tasks involved in data science, you will explore data analytics with Scala in the second module. You'll see how Scala can be used to make sense of data through easy to follow recipes. You will learn about Bokeh bindings for exploratory data analysis and quintessential machine learning with algorithms with Spark ML library. You'll get a sufficient understanding of Spark streaming, machine learning for streaming data, and Spark graphX. Armed with a

firm understanding of data analysis, you will be ready to explore the most cutting-edge aspect of data science — machine learning. The final module teaches you the A to Z of machine learning with Scala. You'll explore Scala for dependency injections and implicits, which are used to write machine learning algorithms. You'll also explore machine learning topics such as clustering, dimentionality reduction, Naive Bayes, Regression models, SVMs, neural networks, and more. This learning path combines some of the best that Packt has to offer into one complete, curated package. It includes content from the following Packt products: Scala for Data Science, Pascal Bugnion Scala Data Analysis Cookbook, Arun Manivannan Scala for Machine Learning, Patrick R. Nicolas Style and approach A complete package with all the information necessary to start building useful data engineering and data science solutions straight away. It contains a diverse set of recipes that cover the full spectrum of interesting data analysis tasks and will help you revolutionize your data analysis skills using Scala.
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.
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.
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.
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.
Troubleshooting Ubuntu Server Skanda Bhargav 2015-09-25 Make life at the office easier for server administrators by helping them build resilient Ubuntu server systems About This Book Tackle the issues you come across in keeping your Ubuntu server up and running Build server machines and troubleshoot cloud computing related issues using Open Stack Discover tips and best practices to be followed for minimum maintenance of Ubuntu Server 3 Who This Book Is For This book is for a vast audience of Linux system administrators who primarily work on Debian-based systems and spend long hours trying fix issues with the enterprise server. Ubuntu is already one of the most popular OSes and this book targets the most common issues that most administrators have to deal with. With the right tools and definite solutions, you will be able to keep your Ubuntu servers in the pink of health. What You Will Learn Deploy packages and their dependencies with repositories Set up your own DNS and network for Ubuntu Server Authenticate and validate users and their access to various systems and services Maintain, monitor, and optimize your server resources and avoid tremendous load Get to know about processes, assigning and changing priorities, and running processes in background Optimize your shell with tools and provide users with an improved shell experience Set up separate environments for various services and run them safely in isolation Understand, build, and deploy OpenStack on your Ubuntu Server In Detail Ubuntu is becoming one of the favorite Linux flavors for many enterprises and is being adopted to a large extent. It supports a wide variety of common network systems and the use of standard Internet services including file serving, e-mail, Web, DNS, and database management. A large scale use and implementation of Ubuntu on servers has given rise to a vast army of Linux administrators who battle it out day in and day out to make sure the systems are in the right frame of operation and pre-empt any untoward incidents that may result in catastrophes for the businesses using it. Despite all these efforts, glitches and bugs occur that affect Ubuntu server's network, memory, application, and hardware and also generate cloud computing related issues using OpenStack. This book will help you end to end. Right from setting up your new Ubuntu Server to learning the best practices to host OpenStack without any hassles. You will be able to control the priority of jobs, restrict or allow access users to certain services, deploy packages, tackle issues related to server effectively, and reduce downtime. Also, you will learn to set up OpenStack, and manage and monitor its services while tuning the machine with best practices. You will also get to know about Virtualization to make services serve users better. Chapter by chapter, you will learn to add new features and functionalities and make your Ubuntu server a full-fledged, production-ready system. Style and approach This book contains topic-by-topic discussion in an easy-to-understand language with loads of examples to help you take care of Ubuntu Server. Plenty of screenshots will guide you through a step-by-step approach.
Applied OpenStack Design Patterns Uchit Vyas 2016-12-20 Learn practical and applied OpenStack cloud design solutions to gain maximum control over your infrastructure. You will achieve a complete controlled and customizable platform. Applied OpenStack Design Patterns teaches you how to map your application flow once you set up components and architectural design patterns. Also covered is storage management and computing to map user requests and allocations. Best practices of High Availability and Native Cluster Management are included. Solutions are presented to network components of OpenStack and to reduce latency and enable faster communication gateways between components of OpenStack and native applications. What You Will Learn: Design a modern cloud infrastructure Solve complex infrastructure application problems Understand OpenStack cloud infrastructure components Adopt a business impact analysis to support existing/new cloud infrastructure Use specific components to integrate an existing tool-chain set to gain agility and a quick, continuous delivery model Who This Book Is For: Seasoned solution architects, DevOps, and system engineers and analysts
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 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
Big Data SMACK Raul Estrada 2016-09-29 Learn how to integrate full-stack open source big data architecture and to choose the correct technology—Scala/Spark, Mesos, Akka, Cassandra, and Kafka—in every layer. Big data architecture is becoming a requirement for many different enterprises. So far, however, the focus has largely been on collecting, aggregating, and crunching large data sets in a timely manner. In many cases now, organizations need more than one paradigm to perform efficient analyses. Big Data SMACK explains each of the full-stack technologies and, more importantly, how to best integrate them. It provides detailed coverage of the practical benefits of these technologies and incorporates real-world examples in every situation. This book focuses on the problems and scenarios solved by the architecture, as well as the solutions provided by every technology. It covers the six main concepts of big data architecture and how integrate, replace, and reinforce every layer: The language: Scala The engine: Spark (SQL, MLib, Streaming, GraphX) The container: Mesos, Docker The view: Akka The storage: Cassandra The message broker: Kafka What You Will Learn: Make big data architecture without using complex Greek letter architectures Build a cheap but effective cluster infrastructure Make queries, reports, and graphs that business demands Manage and exploit unstructured and No-SQL data sources Use tools to monitor the performance of your architecture Integrate all technologies and decide which ones replace and which ones reinforce Who This Book Is For: Developers, data architects, and data scientists looking to integrate the most successful big data open stack architecture and to choose the correct technology in every layer
Learning PySpark Tomasz Drabas 2017-02-27 Build data-intensive applications locally and deploy at scale using the combined powers of Python and Spark 2.0 About This Book Learn why and how you can efficiently use Python to process data and build machine learning models in Apache Spark 2.0 Develop and deploy efficient, scalable real-time Spark solutions Take your understanding of using Spark with Python to the next level with this jump start guide Who This Book Is For If you are a Python developer who wants to learn about the Apache Spark 2.0 ecosystem, this book is for you. A firm understanding of Python is expected to get the best out of the book. Familiarity with Spark would be useful, but is not mandatory. What You Will Learn Learn about Apache Spark and the Spark 2.0 architecture Build and interact with Spark DataFrames using Spark SQL Learn how to solve graph and deep learning problems using GraphFrames and TensorFrames respectively Read, transform, and understand data and use it to train machine learning models Build machine learning models with MLlib and ML Learn how to submit your applications programmatically using spark-submit Deploy locally built applications to a cluster In Detail Apache Spark is an open source framework for efficient cluster computing with a strong interface for data parallelism and fault tolerance. This book will show you how to leverage the power of Python and put it to use in the Spark ecosystem. You will start by getting a firm understanding of the Spark 2.0 architecture and how to set up a Python environment for Spark. You will get familiar with the modules available in PySpark. You will learn how to abstract data with RDDs and DataFrames and understand the streaming capabilities of PySpark. Also, you will get a thorough overview of machine learning capabilities of PySpark using ML and MLlib, graph processing using GraphFrames, and polyglot persistence using Blaze. Finally, you will learn how to deploy your applications to the cloud using the spark-submit command. By the end of this book, you will have established a firm understanding of the Spark Python API and how it can be used to build data-intensive applications. Style and approach This book takes a very comprehensive, step-by-step approach so you understand how the Spark ecosystem can be used with Python to develop efficient, scalable solutions. Every chapter is standalone and written in a very easy-to-understand manner, with a focus on both the hows and the whys of each concept.
Machine Learning with R Cookbook Yu-Wei, Chiu (David Chiu) 2015-03-26 The R language is a powerful open source functional programming language. At its core, R is a statistical programming language that provides impressive tools to analyze data and create high-

level graphics. This book covers the basics of R by setting up a user-friendly programming environment and performing data ETL in R. Data exploration examples are provided that demonstrate how powerful data visualization and machine learning is in discovering hidden relationships. You will then dive into important machine learning topics, including data classification, regression, clustering, association rule mining, and dimension reduction.

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.

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!

MongoDB Cookbook Cyrus Dasadia 2016-01-13 Harness the latest features of MongoDB 3 with this collection of 80 recipes – from managing cloud platforms to app development, this book is a vital resource About This Book Get to grips with the latest features of MongoDB 3 Interact with the MongoDB server and perform a wide range of query operations from the shell From administration to automation, this cookbook keeps you up to date with the world's leading NoSQL database Who This Book Is For This book is engineered for anyone who is interested in managing data in an easy and efficient way using MongoDB. You do not need any prior knowledge of MongoDB, but it would be helpful if you have some programming experience in either Java or Python. What You Will Learn Install, configure, and administer MongoDB sharded clusters and replica sets Begin writing applications using MongoDB in Java and Python languages Initialize the server in three different modes with various configurations Perform cloud deployment and introduce PaaS for Mongo Discover frameworks and products built to improve developer productivity using Mongo Take an in-depth look at the Mongo programming driver APIs in Java and Python Set up enterprise class monitoring and backups of MongoDB In Detail MongoDB is a high-performance and feature-rich NoSQL database that forms the backbone of the systems that power many different organizations – it's easy to see why it's the most popular NoSQL database on the market. Packed with many features that have become essential for many different types of software professionals and incredibly easy to use, this cookbook contains many solutions to the everyday challenges of MongoDB, as well as guidance on effective techniques to extend your skills and capabilities. This book starts with how to initialize the server in three different modes with various configurations. You will then be introduced to programming language drivers in both Java and Python. A new feature in MongoDB 3 is that you can connect to a single node using Python, set to make MongoDB even more popular with anyone working with Python. You will then learn a range of further topics including advanced query operations, monitoring and backup using MMS, as well as some very useful administration recipes including SCRAM-SHA-1 Authentication. Beyond that, you will also find recipes on cloud deployment, including guidance on how to work with Docker containers alongside MongoDB, integrating the database with Hadoop, and tips for improving developer productivity. Created as both an accessible tutorial and an easy to use resource, on hand whenever you need to solve a problem, MongoDB Cookbook will help you handle everything from administration to automation with MongoDB more effectively than ever before. Style and approach Every recipe is explained in a very simple set-by-step manner yet is extremely comprehensive.

Einführung in Apache Solr Markus Klose 2014-02

Ceph Cookbook Karan Singh 2016-02-29 Over 100 effective recipes to help you design, implement, and manage the software-defined and massively scalable Ceph storage system About This Book Implement a Ceph cluster successfully and gain deep insights into its best practices Harness the abilities of experienced storage administrators and architects, and run your own software-defined storage system This comprehensive, step-by-step guide will show you how to build and manage Ceph storage in production environment Who This Book Is For This book is aimed at storage and cloud system engineers, system administrators, and technical architects who are interested in building software-defined storage solutions to power their cloud and 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 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 gain practical tips to 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 Build a Dropbox-like file sync and share service and Ceph federated gateway setup Gain hands-on experience with Calamari 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 system maintenance In Detail Ceph is a unified, distributed storage system designed for excellent performance, reliability, and scalability. This cutting-edge technology has been transforming the storage industry, and is evolving rapidly as a leader in software-defined storage space, extending full support to cloud platforms such as Openstack and Cloudstack, including virtualization platforms. It is the most popular storage backend for Openstack, public, and private clouds, so is the first choice for a storage solution. Ceph is backed by RedHat and is developed by a thriving open source community of individual developers as well as several companies across the globe. This book takes you from a basic knowledge of Ceph to an expert understanding of the most advanced features, walking you through building up a production-grade Ceph storage cluster and helping you develop all the skills you need to plan, deploy, and effectively manage your Ceph cluster. Beginning with the basics, you'll create a Ceph cluster, followed by block, object, and file storage provisioning. Next, you'll get a step-by-step tutorial on integrating it with OpenStack and building a Dropbox-like object storage solution. We'll also take a look at federated architecture and CephFS, and you'll dive into Calamari and VSM for monitoring the Ceph environment. You'll develop expert knowledge on troubleshooting and benchmarking your Ceph storage cluster. Finally, you'll get to grips with the best practices to operate Ceph in a production environment. Style and approach This step-by-step guide is filled with practical tutorials, making complex scenarios easy to understand.

Mastering Large Datasets with Python John Wolohan 2020-01-15 Summary Modern data science solutions need to be clean, easy to read, and scalable. In Mastering Large Datasets with Python, author J.T. Wolohan teaches you how to take a small project and scale it up using a functionally influenced approach to Python coding. You'll explore methods and built-in Python tools that lend themselves to clarity and scalability, like the high-performing parallelism method, as well as distributed technologies that allow for high data throughput. The abundant hands-on exercises in this practical tutorial will lock in these essential skills for any large-scale data science project. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Programming techniques that work well on laptop-sized data can slow to a crawl—or fail altogether—when applied to massive files or distributed datasets. By mastering the powerful map and reduce paradigm, along with the Python-based tools that support it, you can write data-centric applications that scale efficiently without requiring codebase rewrites as your requirements change. About the book Mastering Large Datasets with Python teaches you to write code that can handle datasets of any size. You'll start with laptop-sized datasets that teach you to parallelize data analysis by breaking large tasks into smaller ones that can run simultaneously. You'll then scale those same programs to industrial-sized datasets on a cluster of cloud servers. With the map and reduce paradigm firmly in place, you'll explore tools like Hadoop and PySpark to efficiently process massive distributed datasets, speed up decision-making with machine learning, and simplify your data storage with AWS S3. What's inside An introduction to the map and reduce paradigm Parallelization with the multiprocessing module and pathos framework Hadoop and Spark for distributed computing Running AWS jobs to process large datasets About the reader For Python programmers who need to work faster with more data. About the author J. T. Wolohan is a lead data scientist at Booz Allen Hamilton, and a PhD researcher at Indiana University, Bloomington. Table of Contents: PART 1 1 ¦ Introduction 2 ¦ Accelerating large dataset work: Map and parallel computing 3 ¦ Function pipelines for mapping complex transformations 4 ¦ Processing large datasets with lazy workflows 5 ¦ Accumulation operations with reduce 6 ¦ Speeding up map and reduce with advanced parallelization PART 2 7 ¦ Processing truly big datasets with Hadoop and Spark 8 ¦ Best practices for large data with Apache Streaming and mrjob 9 ¦ PageRank with map and reduce in PySpark 10 ¦ Faster decision-making with machine learning and PySpark PART 3 11 ¦ Large datasets in the cloud with Amazon Web Services and S3 12 ¦ MapReduce in the cloud with Amazon's Elastic MapReduce

Sieben Wochen, sieben Datenbanken Eric Redmond 2012

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.

Hadoop MapReduce v2 Cookbook - Second Edition Thilina Gunarathne 2015-02-25 If you are a Big Data enthusiast and wish to use Hadoop v2 to solve your problems, then this book is for you. This book is for Java programmers with little to moderate knowledge of Hadoop MapReduce. This is also a one-stop reference for developers and system admins who want to quickly get up to speed with using Hadoop v2. It would be helpful to have a basic knowledge of software development using Java and a basic working knowledge of Linux.

R: Recipes for Analysis, Visualization and Machine Learning Viswa Viswanathan 2016-11-24 Get savvy with R language and actualize projects aimed at analysis, visualization and machine learning About This Book Proficiently analyze data and apply machine learning techniques Generate visualizations, develop interactive visualizations and applications to understand various data exploratory functions in R Construct a predictive model by using a variety of machine learning packages Who This Book Is For This Learning Path is ideal for those who have been exposed to R, but have not used it extensively yet. It covers the basics of using R and is written for new and intermediate R users interested in learning. This Learning Path also provides in-depth insights into professional techniques for analysis, visualization, and machine learning with R – it will help you increase your R expertise, regardless of your level of experience. What You Will Learn Get data into your R environment and prepare it for analysis Perform exploratory data analyses and generate meaningful visualizations of the data Generate various plots in R using the basic R plotting techniques Create presentations and learn the basics of creating apps in R for your audience Create and inspect the transaction dataset, performing association analysis with the Apriori algorithm Visualize associations in various graph formats and find frequent itemset using the ECLAT algorithm Build, tune, and evaluate predictive models with different machine learning packages Incorporate R and Hadoop to solve machine learning problems on big data In Detail The R language is a powerful, open source, functional programming language. At its core, R is a statistical programming language that provides impressive tools to analyze data and create high-level graphics. This Learning Path is chock-full of recipes. Literally! It aims to excite you with awesome projects focused on analysis, visualization, and machine learning. We'll start off with data analysis – this will show you ways to use R to generate professional analysis reports. We'll then move on to visualizing our data – this provides you with all the guidance needed to get comfortable with data visualization with R. Finally, we'll move into the world of machine learning – this introduces you to data classification, regression, clustering, association rule mining, and dimension reduction. This Learning Path combines some of the best that Packt has to offer in one complete, curated package. It includes content from the following Packt products: R Data Analysis Cookbook by Viswa Viswanathan and Shanthi Viswanathan R Data Visualization Cookbook by Atmajitsinh Gohil Machine Learning with R Cookbook by Yu-Wei, Chiu (David Chiu) Style and approach This course creates a smooth learning path that will teach you how to analyze data and create stunning visualizations. The step-by-step instructions provided for each recipe in this comprehensive Learning Path will show you how to create machine learning projects with R.

SQL Performance Explained Markus Winand 2012