Practical Hadoop Migration How To Integrate Your Rdbms With The Hadoop Ecosystem And Re Architect Relational Applications To Nosql

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.
Re-architect relational applications to NoSQL, integrate relational database management systems with the Hadoop ecosystem, and transform and migrate relational data to and from Hadoop components. This book covers the best-practice design approaches to re-architecting your relational applications and transforming your relational data to optimize concurrency, security, denormalization, and performance. Winner of IBM’s 2012 Gerstner Award for his implementation of big data and data warehouse initiatives and author of Practical Hadoop Security, author Bhushan Lakhe walks you through the entire transition process. First, he lays out the criteria for deciding what blend of re-architecting, migration, and integration between RDBMS and HDFS best meets your transition objectives. Then he demonstrates how to design your transition model. Lakhe proceeds to cover the selection criteria for ETL tools, the implementation steps for migration with SQOOP- and Flume-based data transfers, and transition optimization techniques for tuning partitions, scheduling aggregations, and redesigning ETL. Finally, he assesses the pros and cons of data lakes and Lambda architecture as integrative solutions and illustrates their implementation with real-world case studies. Hadoop/NoSQL solutions do not offer by default certain relational technology features such as role-based access control, locking for concurrent updates, and various tools for measuring and enhancing performance. Practical Hadoop Migration shows how to use open-source tools to emulate such relational functionalities in Hadoop ecosystem components. What You'll Learn Decide whether you should migrate your relational applications to big data technologies or integrate them Transition your relational applications to Hadoop/NoSQL platforms in terms of logical design and physical implementation Discover RDBMS-to-HDFS integration, data transformation, and optimization techniques Consider when to use Lambda architecture and data lake solutions Select and implement Hadoop-based components and applications to speed transition, optimize integrated performance, and emulate relational functionalities Who This Book Is For Database developers, database administrators, enterprise architects, Hadoop/NoSQL developers, and IT leaders. Its secondary readership is project and program managers and advanced students of database and management information systems.
Summary Hadoop in Practice, Second Edition provides over 100 tested, instantly useful techniques that will help you conquer big data, using Hadoop. This revised new edition covers changes and new features in the Hadoop core architecture, including MapReduce 2. Brand new chapters cover YARN and integrating Kafka, Impala, and Spark SQL with Hadoop. You'll also get new and updated techniques for Flume, Sqoop, and Mahout, all of which have seen major new versions recently. In short, this is the most practical, up-to-date coverage of Hadoop available anywhere. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Book It's always a good time to upgrade your Hadoop skills! Hadoop in Practice, Second Edition provides a collection of 104 tested, instantly useful techniques for analyzing real-time streams, moving data securely, machine learning, managing large-scale clusters, and taming big data using Hadoop. This completely revised edition covers changes and new features in Hadoop core, including MapReduce 2 and

YARN. You'll pick up hands-on best practices for integrating Spark, Kafka, and Impala with Hadoop, and get new and updated techniques for the latest versions of Flume, Sqoop, and Mahout. In short, this is the most practical, up-to-date coverage of Hadoop available. Readers need to know a programming language like Java and have basic familiarity with Hadoop. What's Inside Thoroughly updated for Hadoop 2 How to write YARN applications Integrate real-time technologies like Storm, Impala, and Spark Predictive analytics using Mahout and RR Readers need to know a programming language like Java and have basic familiarity with Hadoop. About the Author Alex Holmes works on tough big-data problems. He is a software engineer, author, speaker, and blogger specializing in large-scale Hadoop projects. Table of Contents PART 1 BACKGROUND AND FUNDAMENTALS Hadoop in a heartbeat Introduction to YARN PART 2 DATA LOGISTICS Data serialization—working with text and beyond Organizing and optimizing data in HDFS Moving data into and out of Hadoop PART 3 BIG DATA PATTERNS Applying MapReduce patterns to big data Utilizing data structures and algorithms at scale Tuning, debugging, and testing PART 4 BEYOND MAPREDUCE SQL on Hadoop Writing a YARN application

Until recently, Hadoop deployments existed on hardware owned and run by organizations. Now, of course, you can acquire the computing resources and network connectivity to run Hadoop clusters in the cloud. But there’s a lot more to deploying Hadoop to the public cloud than simply renting machines. This hands-on guide shows developers and systems administrators familiar with Hadoop how to install, use, and manage cloud-born clusters efficiently. You’ll learn how to architect clusters that work with cloud-provider features—not just to avoid pitfalls, but also to take full advantage of these services. You’ll also compare the Amazon, Google, and Microsoft clouds, and learn how to set up clusters in each of them. Learn how Hadoop clusters run in the cloud, the problems they can help you solve, and their potential drawbacks Examine the common concepts of cloud providers, including compute capabilities, networking and security, and storage Build a functional Hadoop cluster on cloud infrastructure, and learn what the major providers require Explore use cases for high availability, relational data with Hive, and complex analytics with Spark Get patterns and practices for running cloud clusters, from designing for price and security to dealing with maintenance

Practical Hadoop MigrationHow to Integrate Your RDBMS with the Hadoop Ecosystem and Re-Architect Relational Applications to NoSQLApress

Fog computing is rapidly expanding in its applications and capabilities through various parts of society. Utilizing different types of virtualization technologies can push this branch of computing to even greater heights. Fog Computing: Breakthroughs in Research and Practice contains a compendium of the latest academic material on the evolving theory and practice related to fog computing. Including innovative studies on distributed fog computing environments, programming models, and access control mechanisms, this publication is an ideal source for programmers, IT professionals, students, researchers, and engineers.

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

As technology grows more effective and refined, businesses and organizations are increasingly taking advantage by automating processes that were once presided over by human workers. As businesses explore the benefits of machine learning, research is necessary to examine the effects of the integration of technology to human workplaces. Advancing Skill Development for Business Managers in Industry 4.0: Emerging Research and Opportunities is an essential publication that examines Industry 4.0 and the important technological applications that revolutionize and disrupt modern organizations, such as artificial intelligence, machine learning, and programming languages, such as Python, to contextualize big data in business and frame the skills necessary for a high-performing modern workforce. The book provides a conceptual framework, analysis, and discussion of the issues concerning organizational behavior through the lens of organizational culture and emotions. Covering topics that include data-driven organizations, the digital business models, and leadership techniques, this book is ideally designed for managers, executives, IT specialists, computer engineers, data scientists, researchers, academicians, and students.

This book constitutes the refereed proceedings of the 13th International Conference on Practical Applications of Agents and Multi-Agent Systems, PAAMS 2015, held in Salamanca, Spain, in June 2015. The 10 revised full papers and 9 short papers were carefully reviewed and selected from 48 submissions are presented together with 17 demonstrations. The articles report on the application and validation of agent-based models, methods and technologies in a number of key application areas, including: agents and the energy grid, agents and the traffic grid, affective computing and agent development, ambient and contextual agents, social simulation and social networks and other agent-based applications.

LinuxONE is a portfolio of hardware, software, and solutions for an enterprise-grade Linux environment. It has been designed to run more transactions faster and with more security and reliability specifically for the open community. It fully embraces open source-based technology. Two servers are available for LinuxONE: The IBM® LinuxONE III LT1 and IBM LinuxONE III LT2. We describe these servers in "IBM LinuxONE servers" on page 5. Aside from still running SUSE Linux Enterprise Server and Red Hat Enterprise Linux Servers, LinuxONE runs Ubuntu, which is popular on x86 hardware. Ubuntu, which runs the cloud, smartphones, a computer that can remote control a planetary rover for NASA, many market-leading companies, and the Internet of Things, is now available on IBM LinuxONE servers. Together, these two technology communities deliver the perfect environment for cloud and DevOps. Ubuntu 16.04 on LinuxONE offers developers, enterprises, and Cloud Service Providers a scalable and secure

platform for next generation applications that include OpenStack, KVM, Docker, and JuJu. The following are reasons why you would want to optimize your servers through virtualization using LinuxONE: Too many distributed physical servers with low utilization A lengthy provisioning process that delays the implementation of new applications Limitations in data center power and floor space High total cost of ownership (TCO) Difficulty allocating processing power for a dynamic environment This IBM Redbooks® publication provides a technical planning reference for IT organizations that are considering a migration from their x86 distributed servers to LinuxONE. This book walks you through some of the important considerations and planning issues that you might encounter during a migration project. Within the context of a pre-existing UNIX based or x86 environment, it presents an end-to-end view of the technical challenges and methods necessary to complete a successful migration to LinuxONE.

This four volume set LNCS 9528, 9529, 9530 and 9531 constitutes the refereed proceedings of the 15th International Conference on Algorithms and Architectures for Parallel Processing, ICA3PP 2015, held in Zhangjiajie, China, in November 2015. The 219 revised full papers presented together with 77 workshop papers in these four volumes were carefully reviewed and selected from 807 submissions (602 full papers and 205 workshop papers). The first volume comprises the following topics: parallel and distributed architectures; distributed and network-based computing and internet of things and cyber-physical-social computing. The second volume comprises topics such as big data and its applications and parallel and distributed algorithms. The topics of the third volume are: applications of parallel and distributed computing and service dependability and security in distributed and parallel systems. The covered topics of the fourth volume are: software systems and programming models and performance modeling and evaluation.

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

The importance of databases and information systems to the functioning of 21st century life is indisputable. This book presents papers from the 13th International Baltic Conference on Databases and Information Systems, held in Trakai, Lithuania, from 1- 4 July 2018. Since the first of these events in 1994, the Baltic DB&IS has proved itself to be an excellent forum for researchers, practitioners and PhD students to deliver and share their research in the field of advanced information systems, databases and related areas. For the 2018 conference, 69 submissions were received from 15 countries. Each paper was assigned for review to at least three referees from different countries. Following review, 24 regular papers were accepted for presentation at the conference, and from these presented papers the 14 best-revised papers have been selected for publication in this volume, together with a preface and three invited papers written by leading experts. The selected revised and extended papers present original research results in a number of subject areas: information systems, requirements and ontology

engineering; advanced database systems; internet of things; big data analysis; cognitive computing; and applications and case studies. These results will contribute to the further development of this fast-growing field, and will be of interest to all those working with advanced information systems, databases and related areas.

Learn how to form and execute an enterprise information strategy: topics include data governance strategy, data architecture strategy, information security strategy, big data strategy, and cloud strategy. Manage information like a pro, to achieve much better financial results for the enterprise, more efficient processes, and multiple advantages over competitors. As you'll discover in Enterprise Information Management in Practice, EIM deals with both structured data (e.g. sales data and customer data) as well as unstructured data (like customer satisfaction forms, emails, documents, social network sentiments, and so forth). With the deluge of information that enterprises face given their global operations and complex business models, as well as the advent of big data technology, it is not surprising that making sense of the large piles of data is of paramount importance. Enterprises must therefore put much greater emphasis on managing and monetizing both structured and unstructured data. As Saumya Chaki—an information management expert and consultant with IBM—explains in Enterprise Information Management in Practice, it is now more important than ever before to have an enterprise information strategy that covers the entire life cycle of information and its consumption while providing security controls. With Fortune 100 consultant Saumya Chaki as your guide, Enterprise Information Management in Practice covers each of these and the other pillars of EIM in depth, which provide readers with a comprehensive view of the building blocks for EIM. Enterprises today deal with complex business environments where information demands take place in real time, are complex, and often serve as the differentiator among competitors. The effective management of information is thus crucial in managing enterprises. EIM has evolved as a specialized discipline in the business intelligence and enterprise data warehousing space to address the complex needs of information processing and delivery—and to ensure the enterprise is making the most of its information assets.

This three-volume set of books highlights major advances in the development of concepts and techniques in the area of new technologies and architectures of contemporary information systems. Further, it helps readers solve specific research and analytical problems and glean useful knowledge and business value from the data. Each chapter provides an analysis of a specific technical problem, followed by a numerical analysis, simulation and implementation of the solution to the real-life problem. Managing an organisation, especially in today's rapidly changing circumstances, is a very complex process. Increased competition in the marketplace, especially as a result of the massive and successful entry of foreign businesses into domestic markets, changes in consumer behaviour, and broader access to new technologies and information, calls for organisational restructuring and the introduction and modification of management methods using the latest advances in science. This situation has prompted many decision-making bodies to introduce computer modelling of organisation management systems. The three books present the peer-reviewed proceedings of the 39th International Conference "Information Systems Architecture and Technology" (ISAT), held on September 16–18, 2018 in Nysa, Poland. The conference was organised by the Computer Science and Management Systems Departments, Faculty of Computer Science and Management, Wroclaw University of Technology and Sciences and University of Applied Sciences in Nysa, Poland. The papers have been grouped into three major parts: Part I—discusses topics including but not limited to Artificial Intelligence Methods, Knowledge Discovery and Data Mining, Big Data, Knowledge Based Management, Internet of Things, Cloud Computing and High Performance Computing, Distributed Computer Systems, Content Delivery Networks, and Service Oriented Computing. Part II—addresses topics including but not limited to System Modelling for Control, Recognition and Decision Support, Mathematical Modelling in Computer System Design,

Service Oriented Systems and Cloud Computing, and Complex Process Modelling. Part III—focuses on topics including but not limited to Knowledge Based Management, Modelling of Financial and Investment Decisions, Modelling of Managerial Decisions, Production Systems Management and Maintenance, Risk Management, Small Business Management, and Theories and Models of Innovation.

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

Managing Data in Motion describes techniques that have been developed for significantly reducing the complexity of managing system interfaces and enabling scalable architectures. Author April Reeve brings over two decades of experience to present a vendor-neutral approach to moving data between computing environments and systems. Readers will learn the techniques, technologies, and best practices for managing the passage of data between computer systems and integrating disparate data together in an enterprise environment. The average enterprise's computing environment is comprised of hundreds to thousands computer systems that have been built, purchased, and acquired over time. The data from these various systems needs to be integrated for reporting and analysis, shared for business transaction processing, and converted from one format to another when old systems are replaced and new systems are acquired. The management of the "data in motion" in organizations is rapidly becoming one of the biggest concerns for business and IT management. Data warehousing and conversion, real-time data integration, and cloud and "big data" applications are just a few of the challenges facing organizations and businesses today. Managing Data in Motion tackles these and other topics in a style easily understood by business and IT managers as well as programmers and architects. Presents a vendor-neutral overview of the different technologies and techniques for moving data between computer systems including the emerging solutions for unstructured as well as structured data types Explains, in non-technical terms, the architecture and components required to perform data integration Describes how to reduce the complexity of managing system interfaces and enable a scalable data architecture that can handle the dimensions of "Big Data"

Get expert guidance on architecting end-to-end data management solutions with Apache

Hadoop. While many sources explain how to use various components in the Hadoop ecosystem, this practical book takes you through architectural considerations necessary to tie those components together into a complete tailored application, based on your particular use case. To reinforce those lessons, the book's second section provides detailed examples of architectures used in some of the most commonly found Hadoop applications. Whether you're designing a new Hadoop application, or planning to integrate Hadoop into your existing data infrastructure, Hadoop Application Architectures will skillfully guide you through the process. This book covers: Factors to consider when using Hadoop to store and model data Best practices for moving data in and out of the system Data processing frameworks, including MapReduce, Spark, and Hive Common Hadoop processing patterns, such as removing duplicate records and using windowing analytics Giraph, GraphX, and other tools for large graph processing on Hadoop Using workflow orchestration and scheduling tools such as Apache Oozie Near-real-time stream processing with Apache Storm, Apache Spark Streaming, and Apache Flume Architecture examples for clickstream analysis, fraud detection, and data warehousing

This book features research papers presented at the International Conference on Emerging Technologies in Data Mining and Information Security (IEMIS 2020) held at the University of Engineering & Management, Kolkata, India, during July 2020. The book is organized in three volumes and includes high-quality research work by academicians and industrial experts in the field of computing and communication, including full-length papers, research-in-progress papers and case studies related to all the areas of data mining, machine learning, Internet of things (IoT) and information security.

The go-to guidebook for deploying Big Data solutions withHadoop Today's enterprise architects need to understand how the Hadoopframeworks and APIs fit together, and how they can be integrated todeliver real-world solutions. This book is a practical, detailedguide to building and implementing those solutions, with code-levelinstruction in the popular Wrox tradition. It covers storing datawith HDFS and Hbase, processing data with MapReduce, and automatingdata processing with Oozie. Hadoop security, running Hadoop withAmazon Web Services, best practices, and automating Hadoopprocesses in real time are also covered in depth. With in-depth code examples in Java and XML and the latest onrecent additions to the Hadoop ecosystem, this complete resourcealso covers the use of APIs, exposing their inner workings andallowing architects and developers to better leverage and customizethem. The ultimate guide for developers, designers, and architectswho need to build and deploy Hadoop applications Covers storing and processing data with various technologies,automating data processing, Hadoop security, and deliveringreal-time solutions Includes detailed, real-world examples and code-levelguidelines Explains when, why, and how to use these tools effectively Written by a team of Hadoop experts in theprogrammer-to-programmer Wrox style Professional Hadoop Solutions is the reference enterprisearchitects and developers need to maximize the power of Hadoop.

Dive into the world of SQL on Hadoop and get the most out of your Hive data warehouses. This book is your go-to resource for using Hive: authors Scott

Shaw, Ankur Gupta, David Kjerrumgaard, and Andreas Francois Vermeulen take you through learning HiveQL, the SQL-like language specific to Hive, to analyze, export, and massage the data stored across your Hadoop environment. From deploying Hive on your hardware or virtual machine and setting up its initial configuration to learning how Hive interacts with Hadoop, MapReduce, Tez and other big data technologies, Practical Hive gives you a detailed treatment of the software. In addition, this book discusses the value of open source software, Hive performance tuning, and how to leverage semi-structured and unstructured data. What You Will Learn Install and configure Hive for new and existing datasets Perform DDL operations Execute efficient DML operations Use tables, partitions, buckets, and user-defined functions Discover performance tuning tips and Hive best practices Who This Book Is For Developers, companies, and professionals who deal with large amounts of data and could use software that can efficiently manage large volumes of input. It is assumed that readers have the ability to work with SQL.

In this book written for SAP BI, big data, and IT architects, the authors expertly provide clear recommendations for building modern analytics architectures running on SAP HANA technologies. Explore integration with big data frameworks and predictive analytics components. Obtain the tools you need to assess possible architecture scenarios and get guidelines for choosing the best option for your organization. Know your options for on-premise, in the cloud, and hybrid solutions. Readers will be guided through SAP BW/4HANA and SAP HANA native data warehouse scenarios, as well as field-tested integration options with big data platforms. Explore migration options and architecture best practices. Consider organizational and procedural changes resulting from the move to a new, up-to-date analytics architecture that supports your data-driven or data-informed organization. By using practical examples, tips, and screenshots, this book explores: - SAP HANA and SAP BW/4HANA architecture concepts - Predictive Analytics and Big Data component integration - Recommendations for a sustainable, future-proof analytics solutions - Organizational impact and change management

For a large, complex system, the amount of test cases in a regression test suite can range from a few hundred to several thousands, which can take hours or even days to execute. Regression testing also requires considerable resources that are often not readily available. This precludes their use in an interactive setting, further contributing to an inefficient testing process. Cloud computing offers the use of virtualized hardware, effectively unlimited storage, and software services that can help reduce the execution time of large test suites in a cost-effective manner. The research presented by Tilley and Parveen leverages the resources provided by cloud computing infrastructure to facilitate the concurrent execution of test cases. They introduce a decision framework called SMART-T to support migration of software testing to the cloud, a distributed environment called HadoopUnit for the concurrent execution of test cases in the cloud, and a

series of case studies illustrating the use of the framework and the environment. Experimental results indicate a significant reduction in test execution time is possible when compared with a typical sequential environment. Software testing in the cloud is a subject of high interest for advanced practitioners and academic researchers alike. For advanced practitioners, the issue of cloud computing and its impact on the field of software testing is becoming increasingly relevant. For academic researchers, this is a subject that is replete with interesting challenges; there are so many open problems that graduate students will be busy for years to come. To further disseminate results in this field, the authors created a community of interest called “Software Testing in the Cloud” (www.STITC.org), and they encourage all readers to get involved in this exciting new area.
Learn Azure in a Month of Lunches, Second Edition, is a tutorial on writing, deploying, and running applications in Azure. In it, you’ll work through 21 short lessons that give you real-world experience. Each lesson includes a hands-on lab so you can try out and lock in your new skills. Summary You can be incredibly productive with Azure without mastering every feature, function, and service. Learn Azure in a Month of Lunches, Second Edition gets you up and running quickly, teaching you the most important concepts and tasks in 21 practical bite-sized lessons. As you explore the examples, exercises, and labs, you'll pick up valuable skills immediately and take your first steps to Azure mastery! This fully revised new edition covers core changes to the Azure UI, new Azure features, Azure containers, and the upgraded Azure Kubernetes Service. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Microsoft Azure is vast and powerful, offering virtual servers, application templates, and prebuilt services for everything from data storage to AI. To navigate it all, you need a trustworthy guide. In this book, Microsoft engineer and Azure trainer Iain Foulds focuses on core skills for creating cloud-based applications. About the book Learn Azure in a Month of Lunches, Second Edition, is a tutorial on writing, deploying, and running applications in Azure. In it, you’ll work through 21 short lessons that give you real-world experience. Each lesson includes a hands-on lab so you can try out and lock in your new skills. What's inside Understanding Azure beyond point-and-click Securing applications and data Automating your environment Azure services for machine learning, containers, and more About the reader This book is for readers who can write and deploy simple web or client/server applications. About the author Iain Foulds is an engineer and senior content developer with Microsoft. Table of Contents PART 1 - AZURE CORE SERVICES 1 Before you begin 2 Creating a virtual machine 3 Azure Web Apps 4 Introduction to Azure Storage 5 Azure Networking basics PART 2 - HIGH AVAILABILITY AND SCALE 6 Azure Resource Manager 7 High availability and redundancy 8 Load-balancing applications 9 Applications that scale 10 Global databases with Cosmos DB 11 Managing network traffic and routing 12 Monitoring and troubleshooting PART 3 - SECURE BY DEFAULT 13 Backup, recovery, and replication 14 Data encryption

15 Securing information with Azure Key Vault 16 Azure Security Center and updates PART 4 - THE COOL STUFF 17 Machine learning and artificial intelligence 18 Azure Automation 19 Azure containers 20 Azure and the Internet of Things 21 Serverless computing

A Professional Data Engineer authorize data-driven decision making by collecting, transforming, and publishing data. A Data Engineer should be able to blueprint, build, operationalize, secure, and monitor data processing systems with a particular emphasis on security and compliance; scalability and efficiency; reliability and fidelity; and flexibility and portability. A Data Engineer should also be able to leverage, deploy, and continuous train pre-existing machine learning models. Here we've brought best Exam practice questions for Google Cloud so that you can prepare well for Professional Data Engineer exam. Unlike other online simulation practice tests, you get an eBook version that is easy to read & remember these questions. You can simply rely on these questions for successfully certifying this exam.

"This book presents a closer look at the partnership between service oriented architecture and cloud computing environments while analyzing potential solutions to challenges related to the migration of legacy applications"--Provided by publisher.

Congratulations! You completed the MongoDB application within the given tight timeframe and there is a party to celebrate your application's release into production. Although people are congratulating you at the celebration, you are feeling some uneasiness inside. To complete the project on time required making a lot of assumptions about the data, such as what terms meant and how calculations are derived. In addition, the poor documentation about the application will be of limited use to the support team, and not investigating all of the inherent rules in the data may eventually lead to poorly-performing structures in the not-so-distant future. Now, what if you had a time machine and could go back and read this book. You would learn that even NoSQL databases like MongoDB require some level of data modeling. Data modeling is the process of learning about the data, and regardless of technology, this process must be performed for a successful application. You would learn the value of conceptual, logical, and physical data modeling and how each stage increases our knowledge of the data and reduces assumptions and poor design decisions. Read this book to learn how to do data modeling for MongoDB applications, and accomplish these five objectives: Understand how data modeling contributes to the process of learning about the data, and is, therefore, a required technique, even when the resulting database is not relational. That is, NoSQL does not mean NoDataModeling! Know how NoSQL databases differ from traditional relational databases, and where MongoDB fits. Explore each MongoDB object and comprehend how each compares to their data modeling and traditional relational database counterparts, and learn the basics of adding, querying, updating, and deleting data in MongoDB. Practice a streamlined, template-driven

approach to performing conceptual, logical, and physical data modeling. Recognize that data modeling does not always have to lead to traditional data models! Distinguish top-down from bottom-up development approaches and complete a top-down case study which ties all of the modeling techniques together. This book is written for anyone who is working with, or will be working with MongoDB, including business analysts, data modelers, database administrators, developers, project managers, and data scientists. There are three sections: In Section I, Getting Started, we will reveal the power of data modeling and the tight connections to data models that exist when designing any type of database (Chapter 1), compare NoSQL with traditional relational databases and where MongoDB fits (Chapter 2), explore each MongoDB object and comprehend how each compares to their data modeling and traditional relational database counterparts (Chapter 3), and explain the basics of adding, querying, updating, and deleting data in MongoDB (Chapter 4). In Section II, Levels of Granularity, we cover Conceptual Data Modeling (Chapter 5), Logical Data Modeling (Chapter 6), and Physical Data Modeling (Chapter 7). Notice the “ing” at the end of each of these chapters. We focus on the process of building each of these models, which is where we gain essential business knowledge. In Section III, Case Study, we will explain both top down and bottom up development approaches and go through a top down case study where we start with business requirements and end with the MongoDB database. This case study will tie together all of the techniques in the previous seven chapters. Nike Senior Data Architect Ryan Smith wrote the foreword. Key points are included at the end of each chapter as a way to reinforce concepts. In addition, this book is loaded with hands-on exercises, along with their answers provided in Appendix A. Appendix B contains all of the book’s references and Appendix C contains a glossary of the terms used throughout the text.

Work with petabyte-scale datasets while building a collaborative, agile workplace in the process. This practical book is the canonical reference to Google BigQuery, the query engine that lets you conduct interactive analysis of large datasets. BigQuery enables enterprises to efficiently store, query, ingest, and learn from their data in a convenient framework. With this book, you’ll examine how to analyze data at scale to derive insights from large datasets efficiently. Valliappa Lakshmanan, tech lead for Google Cloud Platform, and Jordan Tigani, engineering director for the BigQuery team, provide best practices for modern data warehousing within an autoscaled, serverless public cloud. Whether you want to explore parts of BigQuery you’re not familiar with or prefer to focus on specific tasks, this reference is indispensable.

This book constitutes the refereed proceedings of the workshops and special session co-located with the 17th International Conference on Practical Applications of Agents and Multi-Agent Systems, PAAMS 2019, held in Ávila, Spain, in June 2019. The total of 26 full and 8 short papers presented in this volume were carefully reviewed and selected from 47 submissions. The book also contains extended abstracts of the doctoral consortium contributions. The papers in this volume stem from the following meetings: Workshop on Agents-Based Solutions for Manufacturing and Supply Chain, AMSC; Second International Workshop on Blockchain Technology for Multi-Agent Systems, BTC4MAS; Workshop on MAS

for Complex Networks and Social Computation; CNSC; Workshop on Multi-Agent Based Applications for Energy Markets, Smart Grids and Sustainable Energy Systems, MASGES; Workshop on Smart Cities and Intelligent Agents, SCIA; and Workshop on Swarm Intelligence and Swarm Robotics, SISR; as well as the special session on Software Agents and Virtualization for Internet of Things, SAVIoTS.

Practical Hadoop Security is an excellent resource for administrators planning a production Hadoop deployment who want to secure their Hadoop clusters. A detailed guide to the security options and configuration within Hadoop itself, author Bhushan Lakhe takes you through a comprehensive study of how to implement defined security within a Hadoop cluster in a hands-on way. You will start with a detailed overview of all the security options available for Hadoop, including popular extensions like Kerberos and OpenSSH, and then delve into a hands-on implementation of user security (with illustrated code samples) with both in-the-box features and with security extensions implemented by leading vendors. No security system is complete without a monitoring and tracing facility, so Practical Hadoop Security next steps you through audit logging and monitoring technologies for Hadoop, as well as ready to use implementation and configuration examples--again with illustrated code samples. The book concludes with the most important aspect of Hadoop security – encryption. Both types of encryptions, for data in transit and data at rest, are discussed at length with leading open source projects that integrate directly with Hadoop at no licensing cost. Practical Hadoop Security: Explains importance of security, auditing and encryption within a Hadoop installation Describes how the leading players have incorporated these features within their Hadoop distributions and provided extensions Demonstrates how to set up and use these features to your benefit and make your Hadoop installation secure without impacting performance or ease of use

More than 80% of all data that is collected by organizations is not in a standard relational database. Instead, it is trapped in unstructured documents, social media posts, machine logs, and so on. Many organizations face significant challenges to manage this deluge of unstructured data, such as the following examples: Pinpointing and activating relevant data for large-scale analytics Lacking the fine-grained visibility that is needed to map data to business priorities Removing redundant, obsolete, and trivial (ROT) data Identifying and classifying sensitive data IBM® Spectrum Discover is a modern metadata management software that provides data insight for petabyte-scale file and Object Storage, storage on-premises, and in the cloud. This software enables organizations to make better business decisions and gain and maintain a competitive advantage. IBM Spectrum® Discover provides a rich metadata layer that enables storage administrators, data stewards, and data scientists to efficiently manage, classify, and gain insights from massive amounts of unstructured data. It improves storage economics, helps mitigate risk, and accelerates large-scale analytics to create competitive advantage and speed critical research. This IBM Redbooks® publication presents several use cases that are focused on artificial intelligence (AI) solutions with IBM Spectrum Discover. This book helps storage administrators and technical specialists plan and implement AI solutions by using IBM Spectrum Discover and several other IBM Storage products.

Learn the fundamental aspects of the business statistics, data mining, and machine learning techniques required to understand the huge amount of data generated by your organization. This book explains practical business analytics through examples, covers the steps involved in using it correctly, and shows you the context in which a particular technique does not make sense. Further, Practical Business Analytics using R helps you understand specific issues faced by organizations and how the solutions to these issues can be facilitated by business analytics. This book will discuss and explore the following through examples and case studies: An introduction to R: data management and R functions The architecture, framework, and life cycle of a business analytics project Descriptive analytics using R: descriptive statistics and data cleaning Data mining: classification, association rules, and clustering Predictive analytics:

simple regression, multiple regression, and logistic regression This book includes case studies on important business analytic techniques, such as classification, association, clustering, and regression. The R language is the statistical tool used to demonstrate the concepts throughout the book. What You Will Learn • Write R programs to handle data • Build analytical models and draw useful inferences from them • Discover the basic concepts of data mining and machine learning • Carry out predictive modeling • Define a business issue as an analytical problem Who This Book Is For Beginners who want to understand and learn the fundamentals of analytics using R. Students, managers, executives, strategy and planning professionals, software professionals, and BI/DW professionals.

Cloud computing presents a promising approach for implementing scalable information and communications technology systems for private and public, individual, community, and business use. Achieving Federated and Self-Manageable Cloud Infrastructures: Theory and Practice overviews current developments in cloud computing concepts, architectures, infrastructures and methods, focusing on the needs of small to medium enterprises. The topic of cloud computing is addressed on two levels: the fundamentals of cloud computing and its impact on the IT world; and an analysis of the main issues regarding the cloud federation, autonomic resource management, and efficient market mechanisms, while supplying an overview of the existing solutions able to solve them. This publication is aimed at both enterprise business managers and research and academic audiences alike.

Big Data: A Tutorial-Based Approach explores the tools and techniques used to bring about the marriage of structured and unstructured data. It focuses on Hadoop Distributed Storage and MapReduce Processing by implementing (i) Tools and Techniques of Hadoop Eco System, (ii) Hadoop Distributed File System Infrastructure, and (iii) efficient MapReduce processing. The book includes Use Cases and Tutorials to provide an integrated approach that answers the ‘What’, ‘How’, and ‘Why’ of Big Data. Features Identifies the primary drivers of Big Data Walks readers through the theory, methods and technology of Big Data Explains how to handle the 4 V’s of Big Data in order to extract value for better business decision making Shows how and why data connectors are critical and necessary for Agile text analytics Includes in-depth tutorials to perform necessary set-ups, installation, configuration and execution of important tasks Explains the command line as well as GUI interface to a powerful data exchange tool between Hadoop and legacy r-dbms databases

Here's what you get in this book: - 300 practice questions and answers spanning the breadth of topics under the data science umbrella - Covers statistics, machine learning, SQL, NoSQL, Hadoop and bioinformatics - Emphasis on real-world application with a chapter on Python libraries for machine learning - Focus on the most frequently asked interview questions. Avoid information overload - Compact format: easy to read, easy to carry, so you can study on-the-go Now, you finally have what you need to crush your data science interview, and land that dream job. About The Author Zack Austin has been building large scale enterprise systems for clients in the media, telecom, financial services and publishing since 2001. He is based in New York City.

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

application, and starting point. What You'll Learn Get to know data lake architecture and design principles Implement data capture and streaming strategies Implement data processing strategies in Hadoop Understand the data lake security framework and availability model Who This Book Is For Big data architects and solution architects

A hands-on guide to leveraging NoSQL databases NoSQL databases are an efficient and powerful tool for storing and manipulating vast quantities of data. Most NoSQL databases scale well as data grows. In addition, they are often malleable and flexible enough to accommodate semi-structured and sparse data sets. This comprehensive hands-on guide presents fundamental concepts and practical solutions for getting you ready to use NoSQL databases. Expert author Shashank Tiwari begins with a helpful introduction on the subject of NoSQL, explains its characteristics and typical uses, and looks at where it fits in the application stack. Unique insights help you choose which NoSQL solutions are best for solving your specific data storage needs. Professional NoSQL: Demystifies the concepts that relate to NoSQL databases, including column-family oriented stores, key/value databases, and document databases. Delves into installing and configuring a number of NoSQL products and the Hadoop family of products. Explains ways of storing, accessing, and querying data in NoSQL databases through examples that use MongoDB, HBase, Cassandra, Redis, CouchDB, Google App Engine Datastore and more. Looks at architecture and internals. Provides guidelines for optimal usage, performance tuning, and scalable configurations. Presents a number of tools and utilities relating to NoSQL, distributed platforms, and scalable processing, including Hive, Pig, RRDtool, Nagios, and more.

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

The distributed computing infrastructure known as ‘the Grid’ has undoubtedly been one of the most successful science-oriented large- scale IT projects of the past 20 years. It is now a fully operational international entity, encompassing several hundred computing sites on all continents and giving access to hundreds of thousands of CPU (central processing unit) cores and hundreds of petabytes of storage, all connected by robust national and international scientific networks. It has evolved to become the main computational platform many scientific communities. This book presents lectures from the Enrico Fermi International School of Physics summer school Grid and Cloud computing: Concepts and Practical Applications, held in Varenna, Italy, in July 2014. The school aimed to cover the conceptual and practical aspects of both the Grid and

Cloud computing. The proceedings included here are divided into eight chapters, with chapters 1, 2, 3 and 8 covering general applications of Grid and Cloud computing in various scientific fields, while chapters 4, 5, 6 and 7 discuss specific technical areas of Grid and Cloud structures. The book will be of interest to all those whose work involves the use of the Grid or Cloud computing.

A Professional Cloud Developer builds scalable and highly available applications using Google-recommended practices and tools. This individual has experience with cloud-native applications, developer tools, managed services, and next-generation databases. A Professional Cloud Developer also has proficiency with at least one general-purpose programming language and is skilled at producing meaningful metrics and logs to debug and trace code. Preparing for google cloud professional cloud developer certification to become a Professional Cloud Developer Certified by Google Cloud? Here we have brought best Exam Questions for you so that you can prepare well for this Exam. Unlike other online simulation practice tests, you get an ebook version that is easy to read & remember these questions. You can simply rely on these questions for successfully certifying this exam.

Enterprises are experimenting with using Hadoop to build Big Data Lakes, but many projects are stalling or failing because the approaches that worked at Internet companies have to be adopted for the enterprise. This practical handbook guides managers and IT professionals from the initial research and decision-making process through planning, choosing products, and implementing, maintaining, and governing the modern data lake. You'll explore various approaches to starting and growing a Data Lake, including Data Warehouse off-loading, analytical sandboxes, and "Data Puddles." Author Alex Gorelik shows you methods for setting up different tiers of data, from raw untreated landing areas to carefully managed and summarized data. You'll learn how to enable self-service to help users find, understand, and provision data; how to provide different interfaces to users with different skill levels; and how to do all of that in compliance with enterprise data governance policies.