Modern Data Architecture With Apache Hadoop

Yeah, reviewing a ebook **Modern Data Architecture With Apache Hadoop** could amass your near associates listings. This is just one of the solutions for you to be successful. As understood, talent does not suggest that you have fabulous points.

Comprehending as skillfully as pact even more than new will find the money for each success. next-door to, the broadcast as with ease as sharpness of this Modern Data Architecture With Apache Hadoop can be taken as without difficulty as picked to act.

Apache Hadoop YARN - Arun Murthy 2014-03-14 "This book is a critically needed resource for the newly released Apache Hadoop 2.0, highlighting YARN as the significant breakthrough that broadens Hadoop beyond the MapReduce paradigm." —From the Foreword by Raymie Stata, CEO of Altiscale The Insider's Guide to

Building Distributed, Big Data Applications with Apache Hadoop™ YARN Apache Hadoop is helping drive the Big Data revolution. Now, its data processing has been completely overhauled: Apache Hadoop YARN provides resource management at data center scale and easier ways to create distributed applications

that process petabytes of data. And now in Apache Hadoop™ YARN, two Hadoop technical leaders show you how to develop new applications and adapt existing code to fully leverage these revolutionary advances. YARN project founder Arun Murthy and project lead Vinod Kumar Vavilapalli demonstrate how YARN increases scalability and cluster utilization, enables new programming models and services, and opens new options beyond Java and batch processing. They walk you through the entire YARN project lifecycle, from installation through deployment. You'll find many examples drawn from the authors' cutting-edge experience—first as Hadoop's earliest developers and implementers at Yahoo! and now as Hortonworks developers moving the platform forward and helping customers succeed with it. Coverage includes YARN's goals, design, architecture, and components—how it expands the Apache Hadoop ecosystem Exploring YARN on a single node Administering YARN clusters and Capacity Scheduler Running existing MapReduce applications Developing a large-scale clustered YARN application Discovering new open source frameworks that run under YARN

The REGTECH Book - Janos Barberis 2019-08-06
The Regulatory Technology Handbook The transformational potential of RegTech has been confirmed in recent years with US$1.2 billion invested in start-ups (2017) and an expected additional spending of US$100 billion by 2020. Regulatory technology will not only provide efficiency gains for compliance and reporting functions, it will radically change market structure and supervision. This book, the first of its kind, is providing a comprehensive and invaluable source of information aimed at corporates, regulators, compliance professionals, start-ups and policy makers. The REGTECH Book brings into a single volume the curated industry expertise delivered by subject

matter experts. It serves as a single reference point to understand the RegTech eco-system and its impact on the industry. Readers will learn foundational notions such as: • The economic impact of digitization and datafication of regulation • How new technologies (Artificial Intelligence, Blockchain) are applied to compliance • Business use cases of RegTech for cost-reduction and new product origination • The future regulatory landscape affecting financial institutions, technology companies and other industries Edited by world-class academics and written by compliance professionals, regulators, entrepreneurs and business leaders, the RegTech Book represents an invaluable resource that paves the way for 21st century regulatory innovation.

Spring Data - Mark Pollack 2012-10-12
You can choose several data access frameworks when building Java enterprise applications that work with relational databases. But what about big data? This hands-on introduction shows you how Spring Data makes it relatively easy to build applications across a wide range of new data access technologies such as NoSQL and Hadoop. Through several sample projects, you'll learn how Spring Data provides a consistent programming model that retains NoSQL-specific features and capabilities, and helps you develop Hadoop applications across a wide range of use-cases such as data analysis, event stream processing, and workflow. You'll also discover the features Spring Data adds to Spring's existing JPA and JDBC support for writing RDBMS-based data access layers. Learn about Spring's template helper classes to simplify the use ofdatabase-specific functionality Explore Spring Data's repository abstraction and advanced query functionality Use Spring Data with Redis (key/value store), HBase(column-family), MongoDB (document database), and Neo4j (graph database) Discover the GemFire distributed data grid solution Export Spring Data JPA-managed entities to the Web as

RESTful web services Simplify the development of HBase applications, using a lightweight object-mapping framework Build example big-data pipelines with Spring Batch and Spring Integration

Practical Machine Learning - Sunila Gollapudi 2016-01-30
Tackle the real-world complexities of modern machine learning with innovative, cutting-edge, techniques About This Book Fully-coded working examples using a wide range of machine learning libraries and tools, including Python, R, Julia, and Spark Comprehensive practical solutions taking you into the future of machine learning Go a step further and integrate your machine learning projects with Hadoop Who This Book Is For This book has been created for data scientists who want to see machine learning in action and explore its real-world application. With guidance on everything from the fundamentals of machine learning and predictive analytics to the latest innovations set to lead the big data revolution into the future, this is an unmissable resource for anyone dedicated to tackling current big data challenges. Knowledge of programming (Python and R) and mathematics is advisable if you want to get started immediately. What You Will Learn Implement a wide range of algorithms and techniques for tackling complex data Get to grips with some of the most powerful languages in data science, including R, Python, and Julia Harness the capabilities of Spark and Hadoop to manage and process data successfully Apply the appropriate machine learning technique to address real-world problems Get acquainted with Deep learning and find out how neural networks are being used at the cutting-edge of machine learning Explore the future of machine learning and dive deeper into polyglot persistence, semantic data, and more In Detail Finding meaning in increasingly larger and more complex datasets is a growing demand of the modern world. Machine learning and predictive

analytics have become the most important approaches to uncover data gold mines. Machine learning uses complex algorithms to make improved predictions of outcomes based on historical patterns and the behaviour of data sets. Machine learning can deliver dynamic insights into trends, patterns, and relationships within data, immensely valuable to business growth and development. This book explores an extensive range of machine learning techniques uncovering hidden tricks and tips for several types of data using practical and real-world examples. While machine learning can be highly theoretical, this book offers a refreshing hands-on approach without losing sight of the underlying principles. Inside, a full exploration of the various algorithms gives you high-quality guidance so you can begin to see just how effective machine learning is at tackling contemporary challenges of big data. This is the only book you need to implement a whole suite of open source tools, frameworks, and languages in machine learning. We will cover the leading data science languages, Python and R, and the underrated but powerful Julia, as well as a range of other big data platforms including Spark, Hadoop, and Mahout. Practical Machine Learning is an essential resource for the modern data scientists who want to get to grips with its real-world application. With this book, you will not only learn the fundamentals of machine learning but dive deep into the complexities of real world data before moving on to using Hadoop and its wider ecosystem of tools to process and manage your structured and unstructured data. You will explore different machine learning techniques for both supervised and unsupervised learning; from decision trees to Naive Bayes classifiers and linear and clustering methods, you will learn strategies for a truly advanced approach to the statistical analysis of data. The book also explores the cutting-edge advancements in machine learning, with worked examples and guidance on deep

learning and reinforcement learning, providing you with practical demonstrations and samples that help take the theory–and mystery–out of even the most advanced machine learning methodologies. Style and approach A practical data science tutorial designed to give you an insight into the practical application of machine learning, this book takes you through complex concepts and tasks in an accessible way. Featuring information on a wide range of data science techniques, Practical Machine Learning is a comprehensive data science resource.

The Modern Data Warehouse in Azure - Matt How 2020-06-15
Build a modern data warehouse on Microsoft's Azure Platform that is flexible, adaptable, and fast—fast to snap together, reconfigure, and fast at delivering results to drive good decision making in your business. Gone are the days when data warehousing projects were lumbering dinosaur-style projects that took forever, drained budgets, and produced business intelligence (BI) just in time to tell you what to do 10 years ago. This book will show you how to assemble a data warehouse solution like a jigsaw puzzle by connecting specific Azure technologies that address your own needs and bring value to your business. You will see how to implement a range of architectural patterns using batches, events, and streams for both data lake technology and SQL databases. You will discover how to manage metadata and automation to accelerate the development of your warehouse while establishing resilience at every level. And you will know how to feed downstream analytic solutions such as Power BI and Azure Analysis Services to empower data-driven decision making that drives your business forward toward a pattern of success. This book teaches you how to employ the Azure platform in a strategy to dramatically improve implementation speed and flexibility of data warehousing systems. You will know how to make correct decisions in design, architecture, and

infrastructure such as choosing which type of SQL engine (from at least three options) best meets the needs of your organization. You also will learn about ETL/ELT structure and the vast number of accelerators and patterns that can be used to aid implementation and ensure resilience. Data warehouse developers and architects will find this book a tremendous resource for moving their skills into the future through cloud-based implementations. What You Will LearnChoose the appropriate Azure SQL engine for implementing a given data warehouse Develop smart, reusable ETL/ELT processes that are resilient and easily maintained Automate mundane development tasks through tools such as PowerShell Ensure consistency of data by creating and enforcing data contracts Explore streaming and event-driven architectures for data ingestionCreate advanced staging layers using Azure Data Lake Gen 2 to feed your data warehouse Who This Book Is For Data warehouse or ETL/ELT developers who wish to implement a data warehouse project in the Azure cloud, and developers currently working in on-premise environments who want to move to the cloud, and for developers with Azure experience looking to tighten up their implementation and consolidate their knowledge **Advanced Methodologies and Technologies in Business Operations and Management** - Khosrow-Pour, D.B.A., Mehdi 2018-09-14 Businesses consistently work on new projects, products, and workflows to remain competitive and successful in the modern business environment. To remain zealous, businesses must employ the most effective methods and tools in human resources, project management, and overall business plan execution as competitors work to succeed as well. Advanced Methodologies and Technologies in Business Operations and Management provides emerging research on business tools such as employee engagement, payout policies, and financial investing to promote operational success. While

highlighting the challenges facing modern organizations, readers will learn how corporate social responsibility and utilizing artificial intelligence improve a company's culture and management. This book is an ideal resource for executives and managers, researchers, accountants, and financial investors seeking current research on business operations and management.

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.

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

Handbook of Research on Big Data Storage and Visualization Techniques - Segall, Richard S. 2018-01-05

The digital age has presented an exponential growth in the amount of data available to individuals looking to draw conclusions based on given or collected information across industries. Challenges associated with the analysis, security, sharing, storage, and visualization of large and complex data sets continue to plague data scientists and analysts alike as traditional data processing applications struggle to adequately manage big data. The Handbook of Research on Big Data Storage and Visualization Techniques is a critical scholarly resource that explores big data analytics and technologies and their role in developing a broad understanding of issues pertaining to the use of big data in multidisciplinary fields. Featuring coverage on a broad range of topics, such as architecture patterns, programing systems, and computational energy, this publication is geared towards professionals, researchers, and students seeking current research and application topics on the subject.

Data Lake for Enterprises - Tomcy John 2017-05-31

A practical guide to implementing your enterprise data lake using Lambda Architecture as the base About This Book Build a full-fledged data lake for your organization with popular big data technologies using the Lambda architecture as the base Delve into the big data technologies required to meet modern day business strategies A highly practical guide to implementing enterprise data lakes with lots of examples and real-world use-cases Who This Book Is For Java developers and architects who would like to implement a data lake for their enterprise will find this book useful. If you want to get hands-on experience with the Lambda Architecture and big data technologies by implementing a

practical solution using these technologies, this book will also help you. What You Will Learn Build an enterprise-level data lake using the relevant big data technologies Understand the core of the Lambda architecture and how to apply it in an enterprise Learn the technical details around Sqoop and its functionalities Integrate Kafka with Hadoop components to acquire enterprise data Use flume with streaming technologies for stream-based processing Understand stream- based processing with reference to Apache Spark Streaming Incorporate Hadoop components and know the advantages they provide for enterprise data lakes Build fast, streaming, and high-performance applications using ElasticSearch Make your data ingestion process consistent across various data formats with configurability Process your data to derive intelligence using machine learning algorithms In Detail The term "Data Lake" has recently emerged as a prominent term in the big data industry. Data scientists can make use of it in deriving meaningful insights that can be used by businesses to redefine or transform the way they operate. Lambda architecture is also emerging as one of the very eminent patterns in the big data landscape, as it not only helps to derive useful information from historical data but also correlates real-time data to enable business to take critical decisions. This book tries to bring these two important aspects — data lake and lambda architecture—together. This book is divided into three main sections. The first introduces you to the concept of data lakes, the importance of data lakes in enterprises, and getting you up-to-speed with the Lambda architecture. The second section delves into the principal components of building a data lake using the Lambda architecture. It introduces you to popular big data technologies such as Apache Hadoop, Spark, Sqoop, Flume, and ElasticSearch. The third section is a highly practical demonstration of putting it all together,

and shows you how an enterprise data lake can be implemented, along with several real-world use-cases. It also shows you how other peripheral components can be added to the lake to make it more efficient. By the end of this book, you will be able to choose the right big data technologies using the lambda architectural patterns to build your enterprise data lake. Style and approach The book takes a pragmatic approach, showing ways to leverage big data technologies and lambda architecture to build an enterprise-level data lake.

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

Digital Success: A Holistic Approach to Digital Transformation for Enterprises and Manufacturers - Alasdair Gilchrist 2018-11-04 What is Digital Transformation, why is it so important and why do so many transformation projects fail? More importantly, what can we do to make our transformation initiative succeed? These are a few of the profound questions that we seek to answer in this book. The anomaly between the number of digital transformation being undertaken and the high failure rate may be due to not enough transformation initiatives taking a holistic approach that encompasses people, culture, organisations, processes, business strategy and objectives into the overall mix instead they are focusing largely on the technology. Therefore in this book we contemplate a holistic approach to digital transformation across the entire spectrum of the business from the perspective of an enterprise or manufacturer. Hence we start with people, culture and the need for agile business development when transforming processes, products and services or business models. We examine the how and why we align and tightly couple business objectives to transformation initiatives. Importantly, we examine the need for company-wide collaboration and integration of data, knowledge, processes and systems and the huge benefits initiatives such as data-democratization can deliver. Then we delve deeper into the specific drivers for successful transformation across a very broad range of business functions from the perspective of IT and Operations, Finance, HR, Sales & Marketing, Manufacturing, Inventory, Supply chain and Post-sales service. We also examine

how technology and processes such as the IoT and advanced data analysis have brought about Industry 4.0 and the Smart Factory not through technology upgrades and point solutions but through a holistic approach to digital transformation. A holistic approach to digital transformation, places people, culture, knowledge, capabilities, and decision-making as the pillars supporting the overall organisation's business strategy, objectives and mission values that are built upon a foundation of technology and processes. Successful digital transformation initiatives do not lose their focus on the overarching business strategy and company objectives being the goal, it's not about technology it is about creating new business value at the strategic level. And in this book we will show you how to do it.What is Digital Transformation, why is it so important and why do so many transformation projects fail? More importantly, what can we do to make our transformation initiative succeed? These are a few of the profound questions that we seek to answer in this book. The anomaly between the number of digital transformation being undertaken and the high failure rate may be due to not enough transformation initiatives taking a holistic approach that encompasses people, culture, organisations, processes, business strategy and objectives into the overall mix instead they are focusing largely on the technology. Therefore in this book we contemplate a holistic approach to digital transformation across the entire spectrum of the business from the perspective of an enterprise or manufacturer. Hence we start with people, culture and the need for agile business development when transforming processes, products and services or business models. We examine the how and why we align and tightly couple business objectives to transformation initiatives. Importantly, we examine the need for company-wide collaboration and integration of data, knowledge, processes and systems and the

huge benefits initiatives such as data-democratization can deliver. Then we delve deeper into the specific drivers for successful transformation across a very broad range of business functions from the perspective of IT and Operations, Finance, HR, Sales & Marketing, Manufacturing, Inventory, Supply chain and Post-sales service. We also examine how technology and processes such as the IoT and advanced data analysis have brought about Industry 4.0 and the Smart Factory not through technology upgrades and point solutions but through a holistic approach to digital transformation. A holistic approach to digital transformation, places people, culture, knowledge, capabilities, and decision-making as the pillars supporting the overall organisation's business strategy, objectives and mission values that are built upon a foundation of technology and processes. Successful digital transformation initiatives do not lose their focus on the overarching business strategy and company objectives being the goal, it's not about technology it is about creating new business value at the strategic level. And in this book we will show you how to do it.

Professional Hadoop - Benoy Antony
2016-05-03
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.

Microsoft Certified Azure Data Fundamentals Study Guide - Jake Switzer 2022-04-14

The most authoritative and complete study guide for people beginning to work with data in the Azure cloud In MC Azure Data Fundamentals Study Guide: Exam DP-900, expert Cloud Solution Architect Jake Switzer delivers a hands-on blueprint to acing the DP-900 Azure data certification. The book prepares you for the test – and for a new career in Azure data analytics, architecture, science, and more – with a laser-focus on the job roles and responsibilities of Azure data professionals. You'll receive a foundational knowledge of core data concepts, like relational and non-relational data and transactional and analytical data workloads, while diving deep into every competency

covered on the DP-900 exam. You'll also get: Access to complimentary online study tools, including hundreds of practice exam questions, electronic flashcards, and a searchable glossary Additional prep assistance with access to Sybex's superior interactive online learning environment and test bank Walkthroughs of skills and knowledge that are absolutely necessary for current and aspiring Azure data pros in introductory roles Perfect for anyone just beginning to work with data in the cloud, MC Azure Data Fundamentals Study Guide: Exam DP-900 is a can't-miss resource for anyone prepping for the DP-900 exam or considering a new career working with Azure data.

Modern Big Data Architectures - Dominik Ryzko 2020-04-09
Provides an up-to-date analysis of big data and multi-agent systems The term Big Data refers to the cases, where data sets are too large or too complex for traditional data-processing software. With the spread of new concepts such as Edge Computing or the Internet of Things, production, processing and consumption of this data becomes more and more distributed. As a result, applications increasingly require multiple agents that can work together. A multi-agent system (MAS) is a self-organized computer system that comprises multiple intelligent agents interacting to solve problems that are beyond the capacities of individual agents. Modern Big Data Architectures examines modern concepts and architecture for Big Data processing and analytics. This unique, up-to-date volume provides joint analysis of big data and multi-agent systems, with emphasis on distributed, intelligent processing of very large data sets. Each chapter contains practical examples and detailed solutions suitable for a wide variety of applications. The author, an internationally-recognized expert in Big Data and distributed Artificial Intelligence, demonstrates how base concepts such as agent, actor, and micro-service have reached a point of

convergence—enabling next generation systems to be built by incorporating the best aspects of the field. This book: Illustrates how data sets are produced and how they can be utilized in various areas of industry and science Explains how to apply common computational models and state-of-the-art architectures to process Big Data tasks Discusses current and emerging Big Data applications of Artificial Intelligence Modern Big Data Architectures: A Multi-Agent Systems Perspective is a timely and important resource for data science professionals and students involved in Big Data analytics, and machine and artificial learning.

Virtualizing Hadoop - George Trujillo 2015-07-14
Plan and Implement Hadoop Virtualization for Maximum Performance, Scalability, and Business Agility Enterprises running Hadoop must absorb rapid changes in big data ecosystems, frameworks, products, and workloads. Virtualized approaches can offer important advantages in speed, flexibility, and elasticity. Now, a world-class team of enterprise virtualization and big data experts guide you through the choices, considerations, and tradeoffs surrounding Hadoop virtualization. The authors help you decide whether to virtualize Hadoop, deploy Hadoop in the cloud, or integrate conventional and virtualized approaches in a blended solution. First, Virtualizing Hadoop reviews big data and Hadoop from the standpoint of the virtualization specialist. The authors demystify MapReduce, YARN, and HDFS and guide you through each stage of Hadoop data management. Next, they turn the tables, introducing big data experts to modern virtualization concepts and best practices. Finally, they bring Hadoop and virtualization together, guiding you through the decisions you'll face in planning, deploying, provisioning, and managing virtualized Hadoop. From security to multitenancy to day-to-day management, you'll find reliable answers for

choosing your best Hadoop strategy and executing it. Coverage includes the following: • Reviewing the frameworks, products, distributions, use cases, and roles associated with Hadoop • Understanding YARN resource management, HDFS storage, and I/O • Designing data ingestion, movement, and organization for modern enterprise data platforms • Defining SQL engine strategies to meet strict SLAs • Considering security, data isolation, and scheduling for multitenant environments • Deploying Hadoop as a service in the cloud • Reviewing the essential concepts, capabilities, and terminology of virtualization • Applying current best practices, guidelines, and key metrics for Hadoop virtualization • Managing multiple Hadoop frameworks and products as one unified system • Virtualizing master and worker nodes to maximize availability and performance • Installing and configuring Linux for a Hadoop environment **Big Data Applications in Industry 4.0** - P.

Kaliraj 2022-02-10
Industry 4.0 is the latest technological innovation in manufacturing with the goal to increase productivity in a flexible and efficient manner. Changing the way in which manufacturers operate, this revolutionary transformation is powered by various technology advances including Big Data analytics, Internet of Things (IoT), Artificial Intelligence (AI), and cloud computing. Big Data analytics has been identified as one of the significant components of Industry 4.0, as it provides valuable insights for smart factory management. Big Data and Industry 4.0 have the potential to reduce resource consumption and optimize processes, thereby playing a key role in achieving sustainable development. Big Data Applications in Industry 4.0 covers the recent advancements that have emerged in the field of Big Data and its applications. The book introduces the concepts and advanced tools and technologies for representing and processing Big Data. It also

covers applications of Big Data in such domains as financial services, education, healthcare, biomedical research, logistics, and warehouse management. Researchers, students, scientists, engineers, and statisticians can turn to this book to learn about concepts, technologies, and applications that solve real-world problems. Features An introduction to data science and the types of data analytics methods accessible today An overview of data integration concepts, methodologies, and solutions A general framework of forecasting principles and applications, as well as basic forecasting models including naïve, moving average, and exponential smoothing models A detailed roadmap of the Big Data evolution and its related technological transformation in computing, along with a brief description of related terminologies The application of Industry 4.0 and Big Data in the field of education The features, prospects, and significant role of Big Data in the banking industry, as well as various use cases of Big Data in banking, finance services, and insurance Implementing a Data Lake (DL) in the cloud and the significance of a data lake in decision making

Optimizing Databricks Workloads - Anirudh Kala 2021-12-24
Accelerate computations and make the most of your data effectively and efficiently on Databricks Key Features: Understand Spark optimizations for big data workloads and maximizing performance Build efficient big data engineering pipelines with Databricks and Delta Lake Efficiently manage Spark clusters for big data processing Book Description: Databricks is an industry-leading, cloud-based platform for data analytics, data science, and data engineering supporting thousands of organizations across the world in their data journey. It is a fast, easy, and collaborative Apache Spark-based big data analytics platform for data science and data engineering in the cloud. In Optimizing Databricks Workloads, you

will get started with a brief introduction to Azure Databricks and quickly begin to understand the important optimization techniques. The book covers how to select the optimal Spark cluster configuration for running big data processing and workloads in Databricks, some very useful optimization techniques for Spark DataFrames, best practices for optimizing Delta Lake, and techniques to optimize Spark jobs through Spark core. It contains an opportunity to learn about some of the real-world scenarios where optimizing workloads in Databricks has helped organizations increase performance and save costs across various domains. By the end of this book, you will be prepared with the necessary toolkit to speed up your Spark jobs and process your data more efficiently. What You Will Learn: Get to grips with Spark fundamentals and the Databricks platform Process big data using the Spark DataFrame API with Delta Lake Analyze data using graph processing in Databricks Use MLflow to manage machine learning life cycles in Databricks Find out how to choose the right cluster configuration for your workloads Explore file compaction and clustering methods to tune Delta tables Discover advanced optimization techniques to speed up Spark jobs Who this book is for: This book is for data engineers, data scientists, and cloud architects who have working knowledge of Spark/Databricks and some basic understanding of data engineering principles. Readers will need to have a working knowledge of Python, and some experience of SQL in PySpark and Spark SQL is beneficial.

Big Data Management and Processing - Kuan-Ching Li 2017-05-19
From the Foreword: "Big Data Management and Processing is [a] state-of-the-art book that deals with a wide range of topical themes in the field of Big Data. The book, which probes many issues related to this exciting and rapidly growing field, covers processing, management, analytics, and applications... [It] is a very valuable addition to the literature. It will serve as a source of up-to-

date research in this continuously developing area. The book also provides an opportunity for researchers to explore the use of advanced computing technologies and their impact on enhancing our capabilities to conduct more sophisticated studies." ---Sartaj Sahni, University of Florida, USA "Big Data Management and Processing covers the latest Big Data research results in processing, analytics, management and applications. Both fundamental insights and representative applications are provided. This book is a timely and valuable resource for students, researchers and seasoned practitioners in Big Data fields. --Hai Jin, Huazhong University of Science and Technology, China Big Data Management and Processing explores a range of big data related issues and their impact on the design of new computing systems. The twenty-one chapters were carefully selected and feature contributions from several outstanding researchers. The book endeavors to strike a balance between theoretical and practical coverage of innovative problem solving techniques for a range of platforms. It serves as a repository of paradigms, technologies, and applications that target different facets of big data computing systems. The first part of the book explores energy and resource management issues, as well as legal compliance and quality management for Big Data. It covers In-Memory computing and In-Memory data grids, as well as co-scheduling for high performance computing applications. The second part of the book includes comprehensive coverage of Hadoop and Spark, along with security, privacy, and trust challenges and solutions. The latter part of the book covers mining and clustering in Big Data, and includes applications in genomics, hospital big data processing, and vehicular cloud computing. The book also analyzes funding for Big Data projects.

A Modern Enterprise Architecture Approach - Dr Mehmet Yildiz 2019-10-07
The revised version of this book to provide

essential guidance, compelling ideas, and unique ways to Enterprise Architects so that they can successfully perform complex enterprise modernisation initiatives transforming from chaos to coherence. This is not an ordinary theory book describing Enterprise Architecture in detail. There are myriad of books on the market and in libraries discussing details of enterprise architecture. My aim here is to highlight success factors and reflect lessons learnt from the field within enterprise modernisation and transformation context. As a practising Senior Enterprise Architect, myself, I read hundreds of those books and articles to learn different views. They have been valuable to me to establish my foundations in the earlier phase of my profession. However, what is missing now is a concise guidance book showing Enterprise Architects the novel approaches, insights from the real-life experience and experimentations, and pointing out the differentiating technologies for enterprise modernisation. If only there were such a guide when I started engaging in modernisation and transformation programs. The biggest lesson learned is the business outcome of the enterprise modernisation. What genuinely matters for business is the return on investment of the enterprise architecture and its monetising capabilities. The rest is the theory because nowadays sponsoring executives, due to economic climate, have no interest, attention, or tolerance for non-profitable ventures. I am sorry for disappointing some idealistic Enterprise Architects, but with due respect, it is the reality, and we cannot change it. This book deals with reality rather than theoretical perfection. Anyone against this view on this climate must be coming from another planet. In this concise, uncluttered and easy-to-read book, I attempt to show the significant pain points and valuable considerations for enterprise modernisation using a structured approach and a simple narration especially considering my audience

from non-English speaking backgrounds. The architectural rigour is still essential. We cannot compromise the rigour aiming to the quality of products and services as a target outcome. However, there must be a delicate balance among architectural rigour, business value, and speed to the market. I applied this pragmatic approach to multiple substantial transformation initiatives and complex modernisations programs. The key point is using an incrementally progressing iterative approach to every aspect of modernisation initiatives, including people, processes, tools, and technologies as a whole. Starting with a high-level view of enterprise architecture to set the context, I provided a dozen of distinct chapters to point out and elaborate on the factors which can make a real difference in dealing with complexity and producing excellent modernisation initiatives. As eminent leaders, Enterprise Architects are the critical talents who can undertake this massive mission using their people and technology skills, in addition to many critical attributes such as calm and composed approach. Let's keep in mind that as Enterprise Architects, we are architects, not firefighters! I have full confidence that this book can provide valuable insights and some 'aha' moments for talented architects like yourself to tackle this enormous mission of turning chaos to coherence.
Hadoop Operations - Eric Sammer 2012-09-26 If you've been asked to maintain large and complex Hadoop clusters, this book is a must. Demand for operations-specific material has skyrocketed now that Hadoop is becoming the de facto standard for truly large-scale data processing in the data center. Eric Sammer, Principal Solution Architect at Cloudera, shows you the particulars of running Hadoop in production, from planning, installing, and configuring the system to providing ongoing maintenance. Rather than run through all possible scenarios, this pragmatic operations guide calls out what works, as demonstrated in

critical deployments. Get a high-level overview of HDFS and MapReduce: why they exist and how they work Plan a Hadoop deployment, from hardware and OS selection to network requirements Learn setup and configuration details with a list of critical properties Manage resources by sharing a cluster across multiple groups Get a runbook of the most common cluster maintenance tasks Monitor Hadoop clusters—and learn troubleshooting with the help of real-world war stories Use basic tools and techniques to handle backup and catastrophic failure

Big Data in Psychiatry and Neurology - Ahmed A. Moustafa 2021-06-24

Big Data in Psychiatry and Neurology provides an up-to-date overview of achievements in the field of big data in Psychiatry and Medicine, including applications of big data methods to aging disorders (e.g., Alzheimer's disease and Parkinson's disease), mood disorders (e.g., major depressive disorder), and drug addiction. This book will help researchers, students and clinicians implement new methods for collecting big datasets from various patient populations. Further, it will demonstrate how to use several algorithms and machine learning methods to analyze big datasets, thus providing individualized treatment for psychiatric and neurological patients. As big data analytics is gaining traction in psychiatric research, it is an essential component in providing predictive models for both clinical practice and public health systems. As compared with traditional statistical methods that provide primarily average group-level results, big data analytics allows predictions and stratification of clinical outcomes at an individual subject level. Discusses longitudinal big data and risk factors surrounding the development of psychiatric disorders Analyzes methods in using big data to treat psychiatric and neurological disorders Describes the role machine learning can play in the analysis of big data Demonstrates the

various methods of gathering big data in medicine Reviews how to apply big data to genetics

Modern Big Data Processing with Hadoop - V Naresh Kumar 2018-03-30
A comprehensive guide to design, build and execute effective Big Data strategies using Hadoop Key Features -Get an in-depth view of the Apache Hadoop ecosystem and an overview of the architectural patterns pertaining to the popular Big Data platform -Conquer different data processing and analytics challenges using a multitude of tools such as Apache Spark, Elasticsearch, Tableau and more -A comprehensive, step-by-step guide that will teach you everything you need to know, to be an expert Hadoop Architect Book Description The complex structure of data these days requires sophisticated solutions for data transformation, to make the information more accessible to the users.This book empowers you to build such solutions with relative ease with the help of Apache Hadoop, along with a host of other Big Data tools. This book will give you a complete understanding of the data lifecycle management with Hadoop, followed by modeling of structured and unstructured data in Hadoop. It will also show you how to design real-time streaming pipelines by leveraging tools such as Apache Spark, and build efficient enterprise search solutions using Elasticsearch. You will learn to build enterprise-grade analytics solutions on Hadoop, and how to visualize your data using tools such as Apache Superset. This book also covers techniques for deploying your Big Data solutions on the cloud Apache Ambari, as well as expert techniques for managing and administering your Hadoop cluster. By the end of this book, you will have all the knowledge you need to build expert Big Data systems. What you will learn Build an efficient enterprise Big Data strategy centered around Apache Hadoop Gain a thorough understanding of using Hadoop with various Big Data frameworks such as Apache

Spark, Elasticsearch and more Set up and deploy your Big Data environment on premises or on the cloud with Apache Ambari Design effective streaming data pipelines and build your own enterprise search solutions Utilize the historical data to build your analytics solutions and visualize them using popular tools such as Apache Superset Plan, set up and administer your Hadoop cluster efficiently Who this book is for This book is for Big Data professionals who want to fast-track their career in the Hadoop industry and become an expert Big Data architect. Project managers and mainframe professionals looking forward to build a career in Big Data Hadoop will also find this book to be useful. Some understanding of Hadoop is required to get the best out of this book.

The Enterprise Big Data Lake - Alex Gorelik 2019-02-21
The data lake is a daring new approach for harnessing the power of big data technology and providing convenient self-service capabilities.

But is it right for your company? This book is based on discussions with practitioners and executives from more than a hundred organizations, ranging from data-driven companies such as Google, LinkedIn, and Facebook, to governments and traditional corporate enterprises. You'll learn what a data lake is, why enterprises need one, and how to build one successfully with the best practices in this book. Alex Gorelik, CTO and founder of Waterline Data, explains why old systems and processes can no longer support data needs in the enterprise. Then, in a collection of essays about data lake implementation, you'll examine data lake initiatives, analytic projects, experiences, and best practices from data experts working in various industries. Get a succinct introduction to data warehousing, big data, and data science Learn various paths enterprises take to build a data lake Explore how to build a self-service model and best practices for providing analysts access to the data Use

different methods for architecting your data lake Discover ways to implement a data lake from experts in different industries

HBase in Action - Amandeep Khurana 2012-11-01
Summary HBase in Action has all the knowledge you need to design, build, and run applications using HBase. First, it introduces you to the fundamentals of distributed systems and large scale data handling. Then, you'll explore real-world applications and code samples with just enough theory to understand the practical techniques. You'll see how to build applications with HBase and take advantage of the MapReduce processing framework. And along the way you'll learn patterns and best practices. About the Technology HBase is a NoSQL storage system designed for fast, random access to large volumes of data. It runs on commodity hardware and scales smoothly from modest datasets to billions of rows and millions of columns. About this Book HBase in Action is an experience-driven guide that shows you how to design, build, and run applications using HBase. First, it introduces you to the fundamentals of handling big data. Then, you'll explore HBase with the help of real applications and code samples and with just enough theory to back up the practical techniques. You'll take advantage of the MapReduce processing framework and benefit from seeing HBase best practices in action. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book. What's Inside When and how to use HBase Practical examples Design patterns for scalable data systems Deployment, integration, and design Written for developers and architects familiar with data storage and processing. No prior knowledge of HBase, Hadoop, or MapReduce is required. Table of Contents PART 1 HBASE FUNDAMENTALS Introducing HBase Getting started Distributed HBase, HDFS, and MapReduce PART 2 ADVANCED CONCEPTS

HBase table design Extending HBase with coprocessors Alternative HBase clients PART 3 EXAMPLE APPLICATIONS HBase by example: OpenTSDB Scaling GIS on HBase PART 4 OPERATIONALIZING HBASE Deploying HBase Operations

Encyclopedia of Information Science and Technology, Fourth Edition - Khosrow-Pour, D.B.A., Mehdi 2017-06-20
In recent years, our world has experienced a profound shift and progression in available computing and knowledge sharing innovations. These emerging advancements have developed at a rapid pace, disseminating into and affecting numerous aspects of contemporary society. This has created a pivotal need for an innovative compendium encompassing the latest trends, concepts, and issues surrounding this relevant discipline area. During the past 15 years, the Encyclopedia of Information Science and Technology has become recognized as one of the landmark sources of the latest knowledge and discoveries in this discipline. The Encyclopedia of Information Science and Technology, Fourth Edition is a 10-volume set which includes 705 original and previously unpublished research articles covering a full range of perspectives, applications, and techniques contributed by thousands of experts and researchers from around the globe. This authoritative encyclopedia is an all-encompassing, well-established reference source that is ideally designed to disseminate the most forward-thinking and diverse research findings. With critical perspectives on the impact of information science management and new technologies in modern settings, including but not limited to computer science, education, healthcare, government, engineering, business, and natural and physical sciences, it is a pivotal and relevant source of knowledge that will benefit every professional within the field of information science and technology and is an invaluable addition to every academic and

corporate library.

Modern Big Data Architectures - Dominik Ryzko 2020-03-31
Provides an up-to-date analysis of big data and multi-agent systems The term Big Data refers to the cases, where data sets are too large or too complex for traditional data-processing software. With the spread of new concepts such as Edge Computing or the Internet of Things, production, processing and consumption of this data becomes more and more distributed. As a result, applications increasingly require multiple agents that can work together. A multi-agent system (MAS) is a self-organized computer system that comprises multiple intelligent agents interacting to solve problems that are beyond the capacities of individual agents. Modern Big Data Architectures examines modern concepts and architecture for Big Data processing and analytics. This unique, up-to-date volume provides joint analysis of big data and multi-agent systems, with emphasis on distributed, intelligent processing of very large data sets. Each chapter contains practical examples and detailed solutions suitable for a wide variety of applications. The author, an internationally-recognized expert in Big Data and distributed Artificial Intelligence, demonstrates how base concepts such as agent, actor, and micro-service have reached a point of convergence—enabling next generation systems to be built by incorporating the best aspects of the field. This book: Illustrates how data sets are produced and how they can be utilized in various areas of industry and science Explains how to apply common computational models and state-of-the-art architectures to process Big Data tasks Discusses current and emerging Big Data applications of Artificial Intelligence Modern Big Data Architectures: A Multi-Agent Systems Perspective is a timely and important resource for data science professionals and students involved in Big Data analytics, and machine and artificial learning.

Conquering Big Data with High Performance Computing - Ritu Arora 2016-09-16
This book provides an overview of the resources and research projects that are bringing Big Data and High Performance Computing (HPC) on converging tracks. It demystifies Big Data and HPC for the reader by covering the primary resources, middleware, applications, and tools that enable the usage of HPC platforms for Big Data management and processing.Through interesting use-cases from traditional and non-traditional HPC domains, the book highlights the most critical challenges related to Big Data processing and management, and shows ways to mitigate them using HPC resources. Unlike most books on Big Data, it covers a variety of alternatives to Hadoop, and explains the differences between HPC platforms and Hadoop.Written by professionals and researchers in a range of departments and fields, this book is designed for anyone studying Big Data and its future directions. Those studying HPC will also find the content valuable.

Scalable Big Data Architecture - Bahaaldine Azarmi 2015-12-31
This book highlights the different types of data architecture and illustrates the many possibilities hidden behind the term "Big Data", from the usage of No-SQL databases to the deployment of stream analytics architecture, machine learning, and governance. Scalable Big Data Architecture covers real-world, concrete industry use cases that leverage complex distributed applications , which involve web applications, RESTful API, and high throughput of large amount of data stored in highly scalable No-SQL data stores such as Couchbase and Elasticsearch. This book demonstrates how data processing can be done at scale from the usage of NoSQL datastores to the combination of Big Data distribution. When the data processing is too complex and involves different processing topology like long running jobs, stream

processing, multiple data sources correlation, and machine learning, it's often necessary to delegate the load to Hadoop or Spark and use the No-SQL to serve processed data in real time. This book shows you how to choose a relevant combination of big data technologies available within the Hadoop ecosystem. It focuses on processing long jobs, architecture, stream data patterns, log analysis, and real time analytics. Every pattern is illustrated with practical examples, which use the different open sourceprojects such as Logstash, Spark, Kafka, and so on. Traditional data infrastructures are built for digesting and rendering data synthesis and analytics from large amount of data. This book helps you to understand why you should consider using machine learning algorithms early on in the project, before being overwhelmed by constraints imposed by dealing with the high throughput of Big data. Scalable Big Data Architecture is for developers, data architects, and data scientists looking for a better understanding of how to choose the most relevant pattern for a Big Data project and which tools to integrate into that pattern.

Mobile Big Data - Georgios Skourletopoulos 2017-10-31
This book reports on the latest advances in mobile technologies for collecting, storing and processing mobile big data in connection with wireless communications. It presents novel approaches and applications in which mobile big data is being applied from an engineering standpoint and addresses future theoretical and practical challenges related to the big data field from a mobility perspective. Further, it provides an overview of new methodologies designed to take mobile big data to the Cloud, enable the processing of real-time streaming events on-the-move and enhance the integration of resource availability through the 'Anywhere, Anything, Anytime' paradigm. By providing both academia and industry researchers and professionals with a timely snapshot of emerging mobile big data-

centric systems and highlighting related pitfalls, as well as potential solutions, the book fills an important gap in the literature and fosters the further development in the area of mobile technologies for exploiting mobile big data.

Designing Cloud Data Platforms - Danil Zburivsky 2021-03-17
In Designing Cloud Data Platforms, Danil Zburivsky and Lynda Partner reveal a six-layer approach that increases flexibility and reduces costs. Discover patterns for ingesting data from a variety of sources, then learn to harness pre-built services provided by cloud vendors. Summary Centralized data warehouses, the long-time defacto standard for housing data for analytics, are rapidly giving way to multi-faceted cloud data platforms. Companies that embrace modern cloud data platforms benefit from an integrated view of their business using all of their data and can take advantage of advanced analytic practices to drive predictions and as yet unimagined data services. Designing Cloud Data Platforms is a hands-on guide to envisioning and designing a modern scalable data platform that takes full advantage of the flexibility of the cloud. As you read, you'll learn the core components of a cloud data platform design, along with the role of key technologies like Spark and Kafka Streams. You'll also explore setting up processes to manage cloud-based data, keep it secure, and using advanced analytic and BI tools to analyze it. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Well-designed pipelines, storage systems, and APIs eliminate the complicated scaling and maintenance required with on-prem data centers. Once you learn the patterns for designing cloud data platforms, you'll maximize performance no matter which cloud vendor you use. About the book In Designing Cloud Data Platforms, Danil Zburivsky and Lynda Partner reveal a six-layer approach that increases

flexibility and reduces costs. Discover patterns for ingesting data from a variety of sources, then learn to harness pre-built services provided by cloud vendors. What's inside Best practices for structured and unstructured data sets Cloud-ready machine learning tools Metadata and real-time analytics Defensive architecture, access, and security About the reader For data professionals familiar with the basics of cloud computing, and Hadoop or Spark. About the author Danil Zburivsky has over 10 years of experience designing and supporting large-scale data infrastructure for enterprises across the globe. Lynda Partner is the VP of Analytics-as-a-Service at Pythian, and has been on the business side of data for over 20 years. Table of Contents 1 Introducing the data platform 2 Why a data platform and not just a data warehouse 3 Getting bigger and leveraging the Big 3: Amazon, Microsoft Azure, and Google 4 Getting data into the platform 5 Organizing and processing data 6 Real-time data processing and analytics 7 Metadata layer architecture 8 Schema management 9 Data access and security 10 Fueling business value with data platforms
Trustworthy Computing and Services - Yuyu Yuan 2014-06-26
This book constitutes the refereed proceedings of the International Standard Conference on Trustworthy Distributed Computing and Services, ISCTCS 2013, held in Beijing, China, in November 2013. The 49 revised full papers presented were carefully reviewed and selected from 267 papers. The topics covered are trustworthy infrastructure; security, survivability and fault tolerance; standards, evaluation and certification; trustworthiness of services.
Big Data Analytics - Kim H. Pries 2015-02-05 With this book, managers and decision makers are given the tools to make more informed decisions about big data purchasing initiatives. Big Data Analytics: A Practical Guide for Managers not only supplies descriptions of common tools, but also surveys the various

products and vendors that supply the big data market.Comparing and contrasting the dif
Practical Hive - Scott Shaw 2016-08-27
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.
Microsoft Certified Azure Data Fundamentals (Exam DP-900) Certification Guide - Marcelo Leite 2022-11-25
Learn how to implement successful Azure Data projects and get the skills to clear the DP-900 certification exam with the help of mock tests and self-assessment scenarios for better preparation Key Features Get the knowledge you need to pass the DP-900 exam on your first attempt Gain fundamental knowledge of the core concepts of working with data in Azure cloud data services Learn through a practical approach and test yourself with mock exams at

the end of the book Book Description Passing the DP-900 Microsoft Azure Data Fundamentals exam opens the door to a myriad of opportunities for working with data services in the cloud. But it is not an easy exam and you'll need a guide to set you up for success and prepare you for a career in Microsoft Azure. Absolutely everything you need to pass the DP-900 exam is covered in this concise handbook. After an introductory chapter covering the core terms and concepts, you'll go through the various roles related to working with data in the cloud and learn the similarities and differences between relational and non-relational databases. This foundational knowledge is crucial, as you'll learn how to provision and deploy Azure's relational and non-relational services in detail later in the book. You'll also gain an understanding of how to glean insights with data analytics at both small and large scales, and how to visualize your insights with Power BI. Once you reach the end of the book, you'll be able to test your knowledge with practice tests with detailed explanations of the correct answers. By the end of this book, you will be armed with the knowledge and confidence to not only pass the DP-900 exam but also have a solid foundation from which to embark on a career in Azure data services. What you will learn Explore the concepts of IaaS and PaaS database services on Azure Query, insert, update, and delete relational data using SQL Explore the concepts of data warehouses in Azure Perform data analytics with an Azure Synapse Analytics workspace Upload and retrieve data in Azure Cosmos DB and Azure HDInsight Provision and deploy non-relational data services in Azure Contextualize the knowledge with real-life use cases Test your progress with a mock exam Who this book is for This book is for data engineers, database administrators, or aspiring data professionals getting ready to take the DP-900 exam. It will also be helpful for those looking for a bit of

guidance on how to be better equipped for Azure-related job roles such as Azure database administrator or Azure data engineer. A basic understanding of core data concepts and relational and non-relational data will help you make the most out of this book, but they're not a pre-requisite.

Data Science and Big Data: An Environment of Computational Intelligence - Witold Pedrycz 2017-03-21
This book presents a comprehensive and up-to-date treatise of a range of methodological and algorithmic issues. It also discusses implementations and case studies, identifies the best design practices, and assesses data analytics business models and practices in industry, health care, administration and business.Data science and big data go hand in hand and constitute a rapidly growing area of research and have attracted the attention of industry and business alike. The area itself has opened up promising new directions of fundamental and applied research and has led to interesting applications, especially those addressing the immediate need to deal with large repositories of data and building tangible, user-centric models of relationships in data. Data is the lifeblood of today's knowledge-driven economy.Numerous data science models are oriented towards end users and along with the regular requirements for accuracy (which are present in any modeling), come the requirements for ability to process huge and varying data sets as well as robustness, interpretability, and simplicity (transparency). Computational intelligence with its underlying methodologies and tools helps address data analytics needs.The book is of interest to those researchers and practitioners involved in data science, Internet engineering, computational intelligence, management, operations research, and knowledge-based systems.

Hadoop Application Architectures - Mark Grover 2015-06-30

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

Actionable Insights with Amazon QuickSight - Manos Samatas 2022-01-28
Build interactive dashboards and storytelling reports at scale with the cloud-native BI tool that integrates embedded analytics and ML-powered insights effortlessly Key FeaturesExplore Amazon QuickSight, manage data sources, and build and share dashboardsLearn best practices from an AWS certified big data solutions architect Manage and monitor dashboards using the QuickSight API and other AWS services such as Amazon CloudTrailBook Description Amazon Quicksight is an exciting new visualization that

rivals PowerBI and Tableau, bringing several exciting features to the table – but sadly, there aren't many resources out there that can help you learn the ropes. This book seeks to remedy that with the help of an AWS-certified expert who will help you leverage its full capabilities. After learning QuickSight's fundamental concepts and how to configure data sources, you'll be introduced to the main analysis-building functionality of QuickSight to develop visuals and dashboards, and explore how to develop and share interactive dashboards with parameters and on-screen controls. You'll dive into advanced filtering options with URL actions before learning how to set up alerts and scheduled reports. Next, you'll familiarize yourself with the types of insights before getting to grips with adding ML insights such as forecasting capabilities, analyzing time series data, adding narratives, and outlier detection to your dashboards. You'll also explore patterns to automate operations and look closer into the API

actions that allow us to control settings. Finally, you'll learn advanced topics such as embedded dashboards and multitenancy. By the end of this book, you'll be well-versed with QuickSight's BI and analytics functionalities that will help you create BI apps with ML capabilities. What you will learnUnderstand the wider AWS analytics ecosystem and how QuickSight fits within itSet up and configure data sources with Amazon QuickSightInclude custom controls and add interactivity to your BI application using parametersAdd ML insights such as forecasting, anomaly detection, and narrativesExplore patterns to automate operations using QuickSight APIsCreate interactive dashboards and storytelling with Amazon QuickSightDesign an embedded multi-tenant analytics architectureFocus on data permissions and how to manage Amazon QuickSight operationsWho this book is for This book is for business intelligence (BI) developers and data analysts who are looking to create interactive dashboards

using data from Lake House on AWS with Amazon QuickSight. It will also be useful for anyone who wants to learn Amazon QuickSight in depth using practical, up-to-date examples. You will need to be familiar with general data visualization concepts before you get started with this book, however, no prior experience with Amazon QuickSight is required.

Data Engineering with Apache Spark, Delta Lake, and Lakehouse - Manoj Kukreja 2021-10-22
Understand the complexities of modern-day data engineering platforms and explore strategies to deal with them with the help of use case scenarios led by an industry expert in big data Key Features Become well-versed with the core concepts of Apache Spark and Delta Lake for building data platforms Learn how to ingest, process, and analyze data that can be later used for training machine learning models Understand how to operationalize data models in production using curated data Book Description

In the world of ever-changing data and schemas, it is important to build data pipelines that can auto-adjust to changes. This book will help you build scalable data platforms that managers, data scientists, and data analysts can rely on. Starting with an introduction to data engineering, along with its key concepts and architectures, this book will show you how to use Microsoft Azure Cloud services effectively for data engineering. You'll cover data lake design patterns and the different stages through which the data needs to flow in a typical data lake. Once you've explored the main features of Delta Lake to build data lakes with fast performance and governance in mind, you'll advance to implementing the lambda architecture using Delta Lake. Packed with practical examples and code snippets, this book takes you through real-world examples based on production scenarios faced by the author in his 10 years of experience working with big data. Finally, you'll cover data lake deployment

strategies that play an important role in provisioning the cloud resources and deploying the data pipelines in a repeatable and continuous way. By the end of this data engineering book, you'll know how to effectively deal with ever-changing data and create scalable data pipelines to streamline data science, ML, and artificial intelligence (AI) tasks. What you will learn Discover the challenges you may face in the data engineering world Add ACID transactions to Apache Spark using Delta Lake Understand effective design strategies to build enterprise-grade data lakes Explore architectural and design patterns for building efficient data ingestion pipelines Orchestrate a data pipeline for preprocessing data using Apache Spark and Delta Lake APIs Automate deployment and monitoring of data pipelines in production Get to grips with securing, monitoring, and managing data pipelines models efficiently Who this book is for This book is for aspiring data engineers and data analysts who are new to the world of data engineering and are looking for a practical guide to building scalable data platforms. If you already work with PySpark and want to use Delta Lake for data engineering, you'll find this book useful. Basic knowledge of Python, Spark, and SQL is expected.

Big Data Processing with Apache Spark - Srini Penchikala 2018-03-13
Apache Spark is a popular open-source big-data processing framework thatÕs built around speed, ease of use, and unified distributed computing architecture. Not only it supports developing applications in different languages like Java, Scala, Python, and R, itÕs also hundred times faster in memory and ten times faster even when running on disk compared to traditional data processing frameworks. Whether you are currently working on a big data project or interested in learning more about topics like machine learning, streaming data processing, and graph data analytics, this book is for you. You can learn about Apache Spark and develop

Spark programs for various use cases in big data analytics using the code examples provided. This book covers all the libraries in Spark ecosystem: Spark Core, Spark SQL, Spark Streaming, Spark ML, and Spark GraphX.