Talend For Data Integration Guide Roberto Marchetto

Every enterprise application creates data, whether it consists of log messages, metrics, user activity, or outgoing messages. Moving all this data is just as important as the data itself. With this updated edition, application architects, developers, and production engineers new to the Kafka streaming platform will learn how to handle data in motion. Additional chapters cover Kafka's AdminClient API, transactions, new security features, and tooling changes. Engineers from Confluent and LinkedIn responsible for developing Kafka explain how to deploy production Kafka clusters, write reliable event-driven microservices, and build scalable stream processing applications with this platform. Through detailed examples, you'll learn Kafka's design principles, reliability guarantees, key APIs, and architecture details, including the replication protocol, the controller, and the storage layer. You'll examine: Best practices for deploying and configuring Kafka Kafka producers and consumers for writing and reading messages Patterns and use-case requirements to ensure reliable data delivery Best practices for building data pipelines and applications with Kafka How to perform monitoring, tuning, and maintenance tasks with Kafka in production The most critical metrics among Kafka's operational measurements Kafka's delivery capabilities for stream processing systems

Introductory, theory-practice balanced text teaching the fundamentals of databases to advanced undergraduates or graduate students in information systems or computer science.

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.
Exploit the visualization capabilities of Kibana and build powerful interactive dashboards About This Book Introduction to data-driven architecture and the Elastic stack Build effective dashboards for data visualization and explore datasets with Elastic Graph A comprehensive guide to learning scalable data visualization techniques in Kibana Who This Book Is For If you are a developer, data visualization engineer, or data scientist who wants to get the best of data visualization at scale then this book is perfect for you. A basic understanding of Elasticsearch and Logstash is required to make the best use of this book. What You Will Learn How to create visualizations in Kibana Ingest log data, structure an Elasticsearch cluster, and create visualization assets in Kibana Embed Kibana visualization on web pages Scaffold, develop, and deploy new Kibana & Timelion customizations Build a metrics dashboard in Timelion based on time series data Use the Graph plugin visualization feature and leverage a graph query Create, implement, package, and deploy a new custom plugin Use Prelert to solve anomaly detection challenges In Detail Kibana is an open source data visualization platform that allows you to interact with your data through stunning, powerful graphics. Its simple, browser-based interface enables you to quickly create and share dynamic dashboards that display changes to Elasticsearch queries in real time. In this book, you'll learn how to use the Elastic stack on top of a data architecture to visualize data in real time. All data architectures have different requirements and

expectations when it comes to visualizing the data, whether it's logging analytics, metrics, business analytics, graph analytics, or scaling them as per your business requirements. This book will help you master Elastic visualization tools and adapt them to the requirements of your project. You will start by learning how to use the basic visualization features of Kibana 5. Then you will be shown how to implement a pure metric analytics architecture and visualize it using Timelion, a very recent and trendy feature of the Elastic stack. You will learn how to correlate data using the brand-new Graph visualization and build relationships between documents. Finally, you will be familiarized with the setup of a Kibana development environment so that you can build a custom Kibana plugin. By the end of this book you will have all the information needed to take your Elastic stack skills to a new level of data visualization. Style and approach This book takes a comprehensive, step-by-step approach to working with the visualization aspects of the Elastic stack. Every concept is presented in a very easy-to-follow manner that shows you both the logic and method of implementation. Real world cases are referenced to highlight how each of the key concepts can be put to practical use.

The Only Complete Technical Primer for MDM Planners, Architects, and Implementers Companies moving toward flexible SOA architectures often face difficult information management and integration challenges. The master data they rely on is often stored and managed in ways that are redundant, inconsistent, inaccessible, non-standardized,

and poorly governed. Using Master Data Management (MDM), organizations can regain control of their master data, improve corresponding business processes, and maximize its value in SOA environments. Enterprise Master Data Management provides an authoritative, vendor-independent MDM technical reference for practitioners: architects, technical analysts, consultants, solution designers, and senior IT decisionmakers. Written by the IBM® data management innovators who are pioneering MDM, this book systematically introduces MDM's key concepts and technical themes, explains its business case, and illuminates how it interrelates with and enables SOA. Drawing on their experience with cutting-edge projects, the authors introduce MDM patterns, blueprints, solutions, and best practices published nowhere else—everything you need to establish a consistent, manageable set of master data, and use it for competitive advantage. Coverage includes How MDM and SOA complement each other Using the MDM Reference Architecture to position and design MDM solutions within an enterprise Assessing the value and risks to master data and applying the right security controls Using PIM-MDM and CDI-MDM Solution Blueprints to address industry-specific information management challenges Explaining MDM patterns as enablers to accelerate consistent MDM deployments Incorporating MDM solutions into existing IT landscapes via MDM Integration Blueprints Leveraging master data as an enterprise asset—bringing people, processes, and technology together with MDM and data governance Best practices in MDM deployment, including data

warehouse and SAP integration

This book comprises selected papers from the 16th International Conference on Intelligent Systems Design and Applications (ISDA'16), which was held in Porto, Portugal from December 1 to16, 2016. ISDA 2016 was jointly organized by the Portugual-based Instituto Superior de Engenharia do Porto and the US-based Machine Intelligence Research Labs (MIR Labs) to serve as a forum for the dissemination of state-of-the-art research and development of intelligent systems, intelligent technologies, and applications. The papers included address a wide variety of themes ranging from theories to applications of intelligent systems and computational intelligence area and provide a valuable resource for students and researchers in academia and industry alike.

Imagine what you could do if scalability wasn't a problem. With this hands-on guide, you'll learn how the Cassandra database management system handles hundreds of terabytes of data while remaining highly available across multiple data centers. This third edition—updated for Cassandra 4.0—provides the technical details and practical examples you need to put this database to work in a production environment. Authors Jeff Carpenter and Eben Hewitt demonstrate the advantages of Cassandra's nonrelational design, with special attention to data modeling. If you're a developer, DBA, or application architect looking to solve a database scaling issue or future-proof your application, this guide helps you harness Cassandra's speed and flexibility.

Understand Cassandra's distributed and decentralized structure Use the Cassandra Query Language (CQL) and cqlsh—the CQL shell Create a working data model and compare it with an equivalent relational model Develop sample applications using client drivers for languages including Java, Python, and Node.js Explore cluster topology and learn how nodes exchange data

Data profiling refers to the activity of collecting data about data, i.e., metadata. Most IT professionals and researchers who work with data have engaged in data profiling, at least informally, to understand and explore an unfamiliar dataset or to determine whether a new dataset is appropriate for a particular task at hand. Data profiling results are also important in a variety of other situations, including query optimization, data integration, and data cleaning. Simple metadata are statistics, such as the number of rows and columns, schema and datatype information, the number of distinct values, statistical value distributions, and the number of null or empty values in each column. More complex types of metadata are statements about multiple columns and their correlation, such as candidate keys, functional dependencies, and other types of dependencies. This book provides a classification of the various types of profilable metadata, discusses popular data profiling tasks, and surveys state-of-the-art profiling algorithms. While most of the book focuses on tasks and algorithms for relational data profiling, we also briefly discuss systems and techniques for profiling non-relational data such as graphs and text. We conclude with a discussion of data profiling challenges

and directions for future work in this area.

A key task that any aspiring data-driven organization needs to learn is data wrangling, the process of converting raw data into something truly useful. This practical guide provides business analysts with an overview of various data wrangling techniques and tools, and puts the practice of data wrangling into context by asking, "What are you trying to do and why?" Wrangling data consumes roughly 50-80% of an analyst's time before any kind of analysis is possible. Written by key executives at Trifacta, this book walks you through the wrangling process by exploring several factors—time, granularity, scope, and structure—that you need to consider as you begin to work with data. You'll learn a shared language and a comprehensive understanding of data wrangling, with an emphasis on recent agile analytic processes used by many of today's data-driven organizations. Appreciate the importance—and the satisfaction—of wrangling data the right way. Understand what kind of data is available Choose which data to use and at what level of detail Meaningfully combine multiple sources of data Decide how to distill the results to a size and shape that can drive downstream analysis

After a short description of the key concepts of big data the book explores on the secrecy and security threats posed especially by cloud based data storage. It delivers conceptual frameworks and models along with case studies of recent technology.

Find the right people with the right skills. This book clarifies best practices for creating high-functioning data integration teams, enabling you to understand the skills and requirements, documents, and solutions for planning, designing, and monitoring both one-time migration and daily integration systems. The growth of data is exploding. With multiple sources of information constantly arriving across enterprise systems, combining these systems into a single,

cohesive, and documentable unit has become more important than ever. But the approach toward integration is much different than in other software disciplines, requiring the ability to code, collaborate, and disentangle complex business rules into a scalable model. Data migrations and integrations can be complicated. In many cases, project teams save the actual migration for the last weekend of the project, and any issues can lead to missed deadlines or, at worst, corrupted data that needs to be reconciled post-deployment. This book details how to plan strategically to avoid these last-minute risks as well as how to build the right solutions for future integration projects. What You Will Learn Understand the “language” of integrations and how they relate in terms of priority and ownership Create valuable documents that lead your team from discovery to deployment Research the most important integration tools in the market today Monitor your error logs and see how the output increases the cycle of continuous improvement Market across the enterprise to provide valuable integration solutions Who This Book Is For The executive and integration team leaders who are building the corresponding practice. It is also for integration architects, developers, and business analysts who need additional familiarity with ETL tools, integration processes, and associated project deliverables.
A practical cookbook on building portals with GateIn including user security, gadgets, and every type of portlet possible.

Harness the power of SQL Server 2017 Integration Services to build your data integration solutions with ease About This Book Acquaint yourself with all the newly introduced features in SQL Server 2017 Integration Services Program and extend your packages to enhance their functionality This detailed, step-by-step guide covers everything you need to develop efficient data integration and data transformation solutions for your organization Who This Book Is For

This book is ideal for software engineers, DW/ETL architects, and ETL developers who need to create a new, or enhance an existing, ETL implementation with SQL Server 2017 Integration Services. This book would also be good for individuals who develop ETL solutions that use SSIS and are keen to learn the new features and capabilities in SSIS 2017. What You Will Learn Understand the key components of an ETL solution using SQL Server 2016-2017 Integration Services Design the architecture of a modern ETL solution Have a good knowledge of the new capabilities and features added to Integration Services Implement ETL solutions using Integration Services for both on-premises and Azure data Improve the performance and scalability of an ETL solution Enhance the ETL solution using a custom framework Be able to work on the ETL solution with many other developers and have common design paradigms or techniques Effectively use scripting to solve complex data issues In Detail SQL Server Integration Services is a tool that facilitates data extraction, consolidation, and loading options (ETL), SQL Server coding enhancements, data warehousing, and customizations. With the help of the recipes in this book, you'll gain complete hands-on experience of SSIS 2017 as well as the 2016 new features, design and development improvements including SCD, Tuning, and Customizations. At the start, you'll learn to install and set up SSIS as well other SQL Server resources to make optimal use of this Business Intelligence tools. We'll begin by taking you through the new features in SSIS 2016/2017 and implementing the necessary features to get a modern scalable ETL solution that fits the modern data warehouse. Through the course of chapters, you will learn how to design and build SSIS data warehouses packages using SQL Server Data Tools. Additionally, you'll learn to develop SSIS packages designed to maintain a data warehouse using the Data Flow and other control flow tasks. You'll also be demonstrated

many recipes on cleansing data and how to get the end result after applying different transformations. Some real-world scenarios that you might face are also covered and how to handle various issues that you might face when designing your packages. At the end of this book, you'll get to know all the key concepts to perform data integration and transformation. You'll have explored on-premises Big Data integration processes to create a classic data warehouse, and will know how to extend the toolbox with custom tasks and transforms. Style and approach This cookbook follows a problem-solution approach and tackles all kinds of data integration scenarios by using the capabilities of SQL Server 2016 Integration Services. This book is well supplemented with screenshots, tips, and tricks. Each recipe focuses on a particular task and is written in a very easy-to-follow manner.

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

The most comprehensive Guide yet of Data Integration. Data incorporation includes rolling into one information inhabiting in dissimilar origins and delivering consumers with a united view of those information. This procedure goes important in a diversity of circumstances, that contain either profit-oriented (when 2 alike businesses demand to combine their databases) and methodical (combining study outcomes as of dissimilar bioinformatics depositories, for example) areas. Data incorporation shows with expanding incidence as the capacity and the

demand to share existent information blows up. It has come to be the center of encompassing hypothetical work, and countless open difficulties stay Unsolved. In administration groups, folks often allude to information incorporation as 'enterprise Information Integration' (EII). There has never been a Data Integration Guide like this. It contains 200 answers, much more than you can imagine; comprehensive answers and extensive details and references, with insights that have never before been offered in print. Get the information you need--fast! This all-embracing guide offers a thorough view of key knowledge and detailed insight. This Guide introduces what you want to know about Data Integration. A quick look inside of some of the subjects covered: Hadoop - Commercially supported Hadoop-related products, Directory service - LDAP implementations, Data integration - Data Integration in the Life Sciences, Fabasoft Mindbreeze - Fabasoft Mindbreeze Solutions, Information Sciences Institute - Customers, startups, and spinoffs, Talend - Data management, Infrastructure optimization - Optimization Models, Data virtualization - Reference Books, CASE tool - Environments, Master data management Solutions, Federated database system - Five Level Schema Architecture for FDBSs, E-GIF - Selection of e-GIF specifications has been driven by, and much more...

A comprehensive guide to mastering the most advanced Hadoop 3 concepts Key Features Get to grips with the newly introduced features and capabilities of Hadoop 3 Crunch and process data using MapReduce, YARN, and a host of tools within the Hadoop ecosystem Sharpen your Hadoop skills with real-world case studies and code Book Description Apache Hadoop is one of the most popular big data solutions for distributed storage and for processing large chunks of data. With Hadoop 3, Apache promises to provide a high-performance, more fault-tolerant, and highly efficient big data processing platform, with a focus on improved scalability and

increased efficiency. With this guide, you'll understand advanced concepts of the Hadoop ecosystem tool. You'll learn how Hadoop works internally, study advanced concepts of different ecosystem tools, discover solutions to real-world use cases, and understand how to secure your cluster. It will then walk you through HDFS, YARN, MapReduce, and Hadoop 3 concepts. You'll be able to address common challenges like using Kafka efficiently, designing low latency, reliable message delivery Kafka systems, and handling high data volumes. As you advance, you'll discover how to address major challenges when building an enterprise-grade messaging system, and how to use different stream processing systems along with Kafka to fulfil your enterprise goals. By the end of this book, you'll have a complete understanding of how components in the Hadoop ecosystem are effectively integrated to implement a fast and reliable data pipeline, and you'll be equipped to tackle a range of real-world problems in data pipelines. What you will learn Gain an in-depth understanding of distributed computing using Hadoop 3 Develop enterprise-grade applications using Apache Spark, Flink, and more Build scalable and high-performance Hadoop data pipelines with security, monitoring, and data governance Explore batch data processing patterns and how to model data in Hadoop Master best practices for enterprises using, or planning to use, Hadoop 3 as a data platform Understand security aspects of Hadoop, including authorization and authentication Who this book is for If you want to become a big data professional by mastering the advanced concepts of Hadoop, this book is for you. You'll also find this book useful if you're a Hadoop professional looking to strengthen your knowledge of the Hadoop ecosystem. Fundamental knowledge of the Java programming language and basics of Hadoop is necessary to get started with this book.

Integrating Hadoop leverages the discipline of data integration and applies it to the Hadoop open-source software framework for storing data on clusters of commodity hardware. It is packed with the need-to-know for managers, architects, designers, and developers responsible for populating Hadoop in the enterprise, allowing you to harness big data and do it in such a way that the solution: · Complies with (and even extends) enterprise standards · Integrates seamlessly with the existing information infrastructure · Fills a critical role within enterprise architecture. Integrating Hadoop covers the gamut of the setup, architecture and possibilities for Hadoop in the organization, including: · Supporting an enterprise information strategy · Organizing for a successful Hadoop rollout · Loading and extracting of data in Hadoop · Managing Hadoop data once it's in the cluster · Utilizing Spark, streaming data, and master data in Hadoop processes - examples are provided to reinforce concepts.

This book collects articles presented at the 13th International Conference on Information Technology- New Generations, April, 2016, in Las Vegas, NV USA. It includes over 100 chapters on critical areas of IT including Web Technology, Communications, Security, and Data Mining.

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

Utilize this practical and easy-to-follow guide to modernize traditional enterprise data warehouse and business intelligence environments with next-generation big data technologies. Next-Generation Big Data takes a holistic approach, covering the most important aspects of modern enterprise big data. The book covers not only the main technology stack but also the next-generation tools and applications used for big data warehousing, data warehouse optimization, real-time and batch data ingestion and processing, real-time data visualization, big data governance, data wrangling, big data cloud deployments, and distributed in-memory big data computing. Finally, the book has an extensive and detailed coverage of big data case studies from Navistar, Cerner, British Telecom, Shopzilla, Thomson Reuters, and Mastercard. What You'll Learn Install Apache Kudu, Impala, and Spark to modernize enterprise data warehouse and business intelligence environments, complete with real-world, easy-to-follow examples, and practical advice Integrate HBase, Solr, Oracle, SQL Server, MySQL, Flume, Kafka, HDFS, and Amazon S3 with Apache Kudu, Impala, and Spark Use StreamSets, Talend, Pentaho, and CDAP for real-time and batch data ingestion and processing Utilize Trifacta,

Alteryx, and Datameer for data wrangling and interactive data processing Turbocharge Spark with Alluxio, a distributed in-memory storage platform Deploy big data in the cloud using Cloudera Director Perform real-time data visualization and time series analysis using Zoomdata, Apache Kudu, Impala, and Spark Understand enterprise big data topics such as big data governance, metadata management, data lineage, impact analysis, and policy enforcement, and how to use Cloudera Navigator to perform common data governance tasks Implement big data use cases such as big data warehousing, data warehouse optimization, Internet of Things, real-time data ingestion and analytics, complex event processing, and scalable predictive modeling Study real-world big data case studies from innovative companies, including Navistar, Cerner, British Telecom, Shopzilla, Thomson Reuters, and Mastercard Who This Book Is For BI and big data warehouse professionals interested in gaining practical and real-world insight into next-generation big data processing and analytics using Apache Kudu, Impala, and Spark; and those who want to learn more about other advanced enterprise topics

Pentaho Data Integration CookbookSecond EditionPackt Publishing Ltd

Build a data platform to the industry-leading standards set by Microsoft’s own infrastructure. Summary In Data Engineering on Azure you will learn how to: Pick the right Azure services for different data scenarios Manage data inventory Implement production quality data modeling, analytics, and machine learning workloads Handle data governance Using DevOps to increase reliability Ingesting, storing, and distributing data Apply best practices for compliance and access control Data Engineering on Azure reveals the data management patterns and techniques that support Microsoft’s own massive data infrastructure. Author Vlad Riscutia, a

data engineer at Microsoft, teaches you to bring an engineering rigor to your data platform and ensure that your data prototypes function just as well under the pressures of production. You'll implement common data modeling patterns, stand up cloud-native data platforms on Azure, and get to grips with DevOps for both analytics and machine learning. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Build secure, stable data platforms that can scale to loads of any size. When a project moves from the lab into production, you need confidence that it can stand up to real-world challenges. This book teaches you to design and implement cloud-based data infrastructure that you can easily monitor, scale, and modify. About the book In Data Engineering on Azure you'll learn the skills you need to build and maintain big data platforms in massive enterprises. This invaluable guide includes clear, practical guidance for setting up infrastructure, orchestration, workloads, and governance. As you go, you'll set up efficient machine learning pipelines, and then master time-saving automation and DevOps solutions. The Azure-based examples are easy to reproduce on other cloud platforms. What's inside Data inventory and data governance Assure data quality, compliance, and distribution Build automated pipelines to increase reliability Ingest, store, and distribute data Production-quality data modeling, analytics, and machine learning About the reader For data engineers familiar with cloud computing and DevOps. About the author Vlad Riscutia is a software architect at Microsoft. Table of Contents 1 Introduction PART 1 INFRASTRUCTURE 2 Storage 3 DevOps 4 Orchestration PART 2 WORKLOADS 5 Processing 6 Analytics 7 Machine learning PART 3 GOVERNANCE 8 Metadata 9 Data quality 10 Compliance 11 Distributing data

A practical guide to implementing a scalable and fast state-of-the-art analytical data estate Key

Features Store and analyze data with enterprise-grade security and auditing Perform batch, streaming, and interactive analytics to optimize your big data solutions with ease Develop and run parallel data processing programs using real-world enterprise scenarios Book Description Azure Data Lake, the modern data warehouse architecture, and related data services on Azure enable organizations to build their own customized analytical platform to fit any analytical requirements in terms of volume, speed, and quality. This book is your guide to learning all the features and capabilities of Azure data services for storing, processing, and analyzing data (structured, unstructured, and semi-structured) of any size. You will explore key techniques for ingesting and storing data and perform batch, streaming, and interactive analytics. The book also shows you how to overcome various challenges and complexities relating to productivity and scaling. Next, you will be able to develop and run massive data workloads to perform different actions. Using a cloud-based big data-modern data warehouse-analytics setup, you will also be able to build secure, scalable data estates for enterprises. Finally, you will not only learn how to develop a data warehouse but also understand how to create enterprise-grade security and auditing big data programs. By the end of this Azure book, you will have learned how to develop a powerful and efficient analytical platform to meet enterprise needs. What you will learn Implement data governance with Azure services Use integrated monitoring in the Azure Portal and integrate Azure Data Lake Storage into the Azure Monitor Explore the serverless feature for ad-hoc data discovery, logical data warehousing, and data wrangling Implement networking with Synapse Analytics and Spark pools Create and run Spark jobs with Databricks clusters Implement streaming using Azure Functions, a serverless runtime environment on Azure Explore the predefined ML services in Azure and use them in your app

Who this book is for This book is for data architects, ETL developers, or anyone who wants to get well-versed with Azure data services to implement an analytical data estate for their enterprise. The book will also appeal to data scientists and data analysts who want to explore all the capabilities of Azure data services, which can be used to store, process, and analyze any kind of data. A beginner-level understanding of data analysis and streaming will be required.

Big Data in Radiation Oncology gives readers an in-depth look into how big data is having an impact on the clinical care of cancer patients. While basic principles and key analytical and processing techniques are introduced in the early chapters, the rest of the book turns to clinical applications, in particular for cancer registries, informatics, radiomics, radiogenomics, patient safety and quality of care, patient-reported outcomes, comparative effectiveness, treatment planning, and clinical decision-making. More features of the book are: Offers the first focused treatment of the role of big data in the clinic and its impact on radiation therapy. Covers applications in cancer registry, radiomics, patient safety, quality of care, treatment planning, decision making, and other key areas. Discusses the fundamental principles and techniques for processing and analysis of big data. Address the use of big data in cancer prevention, detection, prognosis, and management. Provides practical guidance on implementation for clinicians and other stakeholders. Dr. Jun Deng is a professor at the Department of Therapeutic Radiology of Yale University School of Medicine and an ABR board certified medical physicist at Yale-New Haven Hospital. He has received numerous honors and awards such as Fellow of Institute of Physics in 2004, AAPM Medical Physics Travel Grant in 2008, ASTRO IGRT Symposium Travel Grant in 2009, AAPM-IPEM Medical Physics Travel Grant in

2011, and Fellow of AAPM in 2013. Lei Xing, Ph.D., is the Jacob Haimson Professor of Medical Physics and Director of Medical Physics Division of Radiation Oncology Department at Stanford University. His research has been focused on inverse treatment planning, tomographic image reconstruction, CT, optical and PET imaging instrumentations, image guided interventions, nanomedicine, and applications of molecular imaging in radiation oncology. Dr. Xing is on the editorial boards of a number of journals in radiation physics and medical imaging, and is recipient of numerous awards, including the American Cancer Society Research Scholar Award, The Whitaker Foundation Grant Award, and a Max Planck Institute Fellowship.

Learn how to take full advantage of Apache Kafka, the distributed, publish-subscribe queue for handling real-time data feeds. With this comprehensive book, you will understand how Kafka works and how it is designed. Authors Neha Narkhede, Gwen Shapira, and Todd Palino show you how to deploy production Kafka clusters; secure, tune, and monitor them; write rock-solid applications that use Kafka; and build scalable stream-processing applications. Learn how Kafka compares to other queues, and where it fits in the big data ecosystem. Dive into Kafka's internal designPick up best practices for developing applications that use Kafka. Understand the best way to deploy Kafka in production monitoring, tuning, and maintenance tasks. Learn how to secure a Kafka cluster.

We seem to be living in the age of A.I. Everywhere you look, companies are touting their most recent A.I., machine learning, and deep learning breakthroughs, even when they are far short of anything that could be touted as a “breakthrough.” “A.I.” has eclipsed “Blockchain” and “Crypto” as the buzzword of today. Indeed, one of the best ways to raise VC funding is to stick

‘AI’ or ‘ML’ at the front of your prospectus and “.ai” at the end of your website. Separating fact from fiction is more important than it has ever been. The A.I. Marketer breaks down A.I., machine learning, and deep learning into five unique use cases—sound, time series, text, image, and video—and also reveals how marketing executives can utilize this powerful technology to help them more finely tune their marketing campaigns, better segment their customers, increase lead generation, and foster strong customer loyalty. Today, “Personalization”—the process of utilizing mobile, social, geo-location data, web morphing, context and even affective computing to tailor messages and experiences to an individual interacting with them—is becoming the optimum word in a radically new customer intelligence environment. The A.I. Marketer explains this complex technology in simple to understand terms and then shows how marketers can utilize the psychology of personalization with A.I. to both create more effective marketing campaigns as well as increase customer loyalty. Pearson shows companies how to avoid Adobe’s warning of not using industrial-age technology in the digital era. Pearson also reveals how to create a platform of technology that seamlessly integrates EDW and real-time streaming data with social media content. Analytical models and neural nets can then be built on both commerical and open source technology to better understand the customer, thereby strengthening the brand and, just as importantly, increasing ROI.

Would you like to use a consistent visual notation for drawing integration solutions? "Look inside the front cover." Do you want to harness the power of asynchronous systems without getting caught in the pitfalls? "See "Thinking Asynchronously" in the Introduction." Do you want to know which style of application integration is best for your purposes? "See Chapter 2,

Integration Styles." Do you want to learn techniques for processing messages concurrently? "See Chapter 10, Competing Consumers and Message Dispatcher." Do you want to learn how you can track asynchronous messages as they flow across distributed systems? "See Chapter 11, Message History and Message Store." Do you want to understand how a system designed using integration patterns can be implemented using Java Web services, .NET message queuing, and a TIBCO-based publish-subscribe architecture? "See Chapter 9, Interlude: Composed Messaging." Utilizing years of practical experience, seasoned experts Gregor Hohpe and Bobby Woolf show how asynchronous messaging has proven to be the best strategy for enterprise integration success. However, building and deploying messaging solutions presents a number of problems for developers. " Enterprise Integration Patterns " provides an invaluable catalog of sixty-five patterns, with real-world solutions that demonstrate the formidable of messaging and help you to design effective messaging solutions for your enterprise. The authors also include examples covering a variety of different integration technologies, such as JMS, MSMQ, TIBCO ActiveEnterprise, Microsoft BizTalk, SOAP, and XSL. A case study describing a bond trading system illustrates the patterns in practice, and the book offers a look at emerging standards, as well as insights into what the future of enterprise integration might hold. This book provides a consistent vocabulary and visual notation framework to describe large-scale integration solutions across many technologies. It also explores in detail the advantages and limitations of asynchronous messaging architectures. The authors present practical advice on designing code that connects an application to a messaging system, and provide extensive information to help you determine when to send a message, how to route it to the proper destination, and how to monitor the health of a

messaging system. If you want to know how to manage, monitor, and maintain a messaging system once it is in use, get this book. 0321200683B09122003

Learn how to transition from Excel-based business intelligence (BI) analysis to enterprise stacks of open-source BI tools. Select and implement the best free and freemium open-source BI tools for your company's needs and design, implement, and integrate BI automation across the full stack using agile methodologies. Business Intelligence Tools for Small Companies provides hands-on demonstrations of open-source tools suitable for the BI requirements of small businesses. The authors draw on their deep experience as BI consultants, developers, and administrators to guide you through the extract-transform-load/data warehousing (ETL/DWH) sequence of extracting data from an enterprise resource planning (ERP) database freely available on the Internet, transforming the data, manipulating them, and loading them into a relational database. The authors demonstrate how to extract, report, and dashboard key performance indicators (KPIs) in a visually appealing format from the relational database management system (RDBMS). They model the selection and implementation of free and freemium tools such as Pentaho Data Integrator and Talend for ELT, Oracle XE and MySQL/MariaDB for RDBMS, and Qliksense, Power BI, and MicroStrategy Desktop for reporting. This richly illustrated guide models the deployment of a small company BI stack on an inexpensive cloud platform such as AWS. What You'll Learn You will learn how to manage, integrate, and automate the processes of BI by selecting and implementing tools to: Implement and manage the business intelligence/data warehousing (BI/DWH) infrastructure Extract data from any enterprise resource planning (ERP) tool Process and integrate BI data using open-source extract-transform-load (ETL) tools Query, report, and analyze BI data using open-

source visualization and dashboard tools Use a MOLAP tool to define next year's budget, integrating real data with target scenarios Deploy BI solutions and big data experiments inexpensively on cloud platforms Who This Book Is For Engineers, DBAs, analysts, consultants, and managers at small companies with limited resources but whose BI requirements have outgrown the limitations of Excel spreadsheets; personnel in mid-sized companies with established BI systems who are exploring technological updates and more cost-efficient solutions

The proposed book will discuss various aspects of big data Analytics. It will deliberate upon the tools, technology, applications, use cases and research directions in the field. Chapters would be contributed by researchers, scientist and practitioners from various reputed universities and organizations for the benefit of readers.

Data is bigger, arrives faster, and comes in a variety of formats—and it all needs to be processed at scale for analytics or machine learning. But how can you process such varied workloads efficiently? Enter Apache Spark. Updated to include Spark 3.0, this second edition shows data engineers and data scientists why structure and unification in Spark matters. Specifically, this book explains how to perform simple and complex data analytics and employ machine learning algorithms. Through step-by-step walk-throughs, code snippets, and notebooks, you'll be able to: Learn Python, SQL, Scala, or Java high-level Structured APIs Understand Spark operations and SQL Engine Inspect, tune, and debug Spark operations with Spark configurations and Spark UI Connect to data sources: JSON, Parquet, CSV, Avro, ORC, Hive, S3, or Kafka Perform analytics on batch and streaming data using Structured Streaming Build reliable data pipelines with open source Delta Lake and Spark Develop machine learning

pipelines with MLlib and productionize models using MLflow

Migrate your data to Salesforce and build low-maintenance and high-performing data integrations to get the most out of Salesforce and make it a "go-to" place for all your organization's customer information. When companies choose to roll out Salesforce, users expect it to be the place to find any and all Information related to a customer—the coveted Client 360° view. On the day you go live, users expect to see all their accounts, contacts, and historical data in the system. They also expect that data entered in other systems will be exposed in Salesforce automatically and in a timely manner. This book shows you how to migrate all your legacy data to Salesforce and then design integrations to your organization's mission-critical systems. As the Salesforce platform grows more powerful, it also grows in complexity. Whether you are migrating data to Salesforce, or integrating with Salesforce, it is important to understand how these complexities need to be reflected in your design. Developing Data Migrations and Integrations with Salesforce covers everything you need to know to migrate your data to Salesforce the right way, and how to design low-maintenance, high-performing data integrations with Salesforce. This book is written by a practicing Salesforce integration architect with dozens of Salesforce projects under his belt. The patterns and practices covered in this book are the results of the lessons learned during those projects. What You'll Learn Know how Salesforce's data engine is architected and why Use the Salesforce Data APIs to load and extract data Plan and execute your data migration to Salesforce Design low-maintenance, high-performing data integrations with Salesforce Understand common data integration patterns and the pros and cons of each Know real-time integration options for Salesforce Be aware of common pitfalls Build reusable transformation

code covering commonly needed Salesforce transformation patterns Who This Book Is For Those tasked with migrating data to Salesforce or building ongoing data integrations with Salesforce, regardless of the ETL tool or middleware chosen; project sponsors or managers nervous about data tracks putting their projects at risk; aspiring Salesforce integration and/or migration specialists; Salesforce developers or architects looking to expand their skills and take on new challenges

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

The LNCS journal Transactions on Large-Scale Data- and Knowledge-Centered Systems focuses on data management, knowledge discovery, and knowledge processing, which are core and hot topics in computer science. Since the 1990s, the Internet has become the main driving force behind application development in all domains. An increase in the demand for resource sharing (e.g., computing resources, services, metadata, data sources) across different sites connected through networks has led to an evolution of data- and knowledge management systems from centralized systems to decentralized systems enabling large-scale distributed applications providing high scalability. This, the 48th issue of Transactions on Large-Scale Data- and Knowledge-Centered Systems, contains 8 invited papers dedicated to the memory of Prof. Dr. Roland Wagner. The topics covered include distributed database systems, NewSQL, scalable transaction management, strong consistency, caches, data warehouse, ETL, reinforcement learning, stochastic approximation, multi-agent systems, ontology, model-driven development, organisational modelling, digital government, new institutional economics and data governance.

A beginner's guide to analyzing and visualizing your Elasticsearch data using Kibana 7 and Timelion Key Features Gain a fundamental understanding of how Kibana operates within the Elastic Stack Explore your data with Elastic Graph and create rich dashboards in Kibana Learn scalable data visualization techniques in Kibana 7 Book Description Kibana is a window into the Elastic Stack, that enables the visual exploration and real-time analysis of your data in Elasticsearch. This book will help you understand the core concepts of the use of Kibana 7 for rich analytics and data visualization. If you’re new to the tool or want to get to grips with the latest features introduced in Kibana 7, this book is the perfect beginner's guide. You’ll learn

how to set up and configure the Elastic Stack and understand where Kibana sits within the architecture. As you advance, you'll learn how to ingest data from different sources using Beats or Logstash into Elasticsearch, followed by exploring and visualizing data in Kibana. Whether working with time series data to create complex graphs using Timelion or embedding visualizations created in Kibana into your web applications, this book covers it all. It also covers topics that every Elastic developer needs to be aware of, such as installing and configuring Application Performance Monitoring (APM) servers and agents. Finally, you'll also learn how to create effective machine learning jobs in Kibana to find anomalies in your data. By the end of this book, you'll have a solid understanding of Kibana, and be able to create your own visual analytics solutions from scratch. What you will learn Explore the data-driven architecture of the Elastic Stack Install and set up Kibana 7 and other Elastic Stack components Use Beats and Logstash to get input from different data sources Create different visualizations using Kibana Build enterprise-grade Elastic dashboards from scratch Use Timelion to play with time series data Install and configure APM servers and APM agents Work with Dev Tools, Spaces, Graph, and other important tools Who this book is for If you're an aspiring Elastic developer or data analysts, this book is for you. You'll also find it useful if you want to get up to speed with the new features of Kibana 7 and perform data visualization on enterprise data. No prior knowledge of Kibana is expected, but some experience with Elasticsearch will be helpful.

Defining a set of guiding principles for data management and describing how these principles can be applied within data management functional areas; Providing a functional framework for the implementation of enterprise data management practices; including widely adopted

practices, methods and techniques, functions, roles, deliverables and metrics; Establishing a common vocabulary for data management concepts and serving as the basis for best practices for data management professionals. DAMA-DMBOK2 provides data management and IT professionals, executives, knowledge workers, educators, and researchers with a framework to manage their data and mature their information infrastructure, based on these principles: Data is an asset with unique properties; The value of data can be and should be expressed in economic terms; Managing data means managing the quality of data; It takes metadata to manage data; It takes planning to manage data; Data management is cross-functional and requires a range of skills and expertise; Data management requires an enterprise perspective; Data management must account for a range of perspectives; Data management is data lifecycle management; Different types of data have different lifecycle requirements; Managing data includes managing risks associated with data; Data management requirements must drive information technology decisions; Effective data management requires leadership commitment. Explore the modern market of data analytics platforms and the benefits of using Snowflake computing, the data warehouse built for the cloud. With the rise of cloud technologies, organizations prefer to deploy their analytics using cloud providers such as Amazon Web Services (AWS), Microsoft Azure, or Google Cloud Platform. Cloud vendors are offering modern data platforms for building cloud analytics solutions to collect data and consolidate into single storage solutions that provide insights for business users. The core of any analytics framework is the data warehouse, and previously customers did not have many choices of platform to use. Snowflake was built specifically for the cloud and it is a true game changer for the analytics market. This book will help onboard you to Snowflake, present best practices to

deploy, and use the Snowflake data warehouse. In addition, it covers modern analytics architecture and use cases. It provides use cases of integration with leading analytics software such as Matillion ETL, Tableau, and Databricks. Finally, it covers migration scenarios for on-premise legacy data warehouses. What You Will Learn Know the key functionalities of Snowflake Set up security and access with cluster Bulk load data into Snowflake using the COPY command Migrate from a legacy data warehouse to Snowflake integrate the Snowflake data platform with modern business intelligence (BI) and data integration tools Who This Book Is For Those working with data warehouse and business intelligence (BI) technologies, and existing and potential Snowflake users

Big data is currently one of the most critical emerging technologies. Organizations around the world are looking to exploit the explosive growth of data to unlock previously hidden insights in the hope of creating new revenue streams, gaining operational efficiencies, and obtaining greater understanding of customer needs. It is important to think of big data and analytics together. Big data is the term used to describe the recent explosion of different types of data from disparate sources. Analytics is about examining data to derive interesting and relevant trends and patterns, which can be used to inform decisions, optimize processes, and even drive new business models. With today's deluge of data comes the problems of processing that data, obtaining the correct skills to manage and analyze that data, and establishing rules to govern the data's use and distribution. The big data technology stack is ever growing and sometimes confusing, even more so when we add the complexities of setting up big data environments with large up-front investments. Cloud computing seems to be a perfect vehicle for hosting big data workloads. However, working on big data in the cloud brings its own

challenge of reconciling two contradictory design principles. Cloud computing is based on the concepts of consolidation and resource pooling, but big data systems (such as Hadoop) are built on the shared nothing principle, where each node is independent and self-sufficient. A solution architecture that can allow these mutually exclusive principles to coexist is required to truly exploit the elasticity and ease-of-use of cloud computing for big data environments. This IBM® RedpaperTM publication is aimed at chief architects, line-of-business executives, and CIOs to provide an understanding of the cloud-related challenges they face and give prescriptive guidance for how to realize the benefits of big data solutions quickly and cost-effectively.

Christian Fürber investigates the useful application of semantic technologies for the area of data quality management. Based on a literature analysis of typical data quality problems and typical activities of data quality management processes, he develops the Semantic Data Quality Management framework as the major contribution of this thesis. The SDQM framework consists of three components that are evaluated in two different use cases. Moreover, this thesis compares the framework to conventional data quality software. Besides the framework, this thesis delivers important theoretical findings, namely a comprehensive typology of data quality problems, ten generic data requirement types, a requirement-centric data quality management process, and an analysis of related work.

This book is written in a concise and easytounderstand manner, and acts as a comprehensive guide on data analytics and integration with Talend big data processing jobs.If you are a chief information officer, enterprise architect, data architect, data scientist, software developer, software engineer, or a data analyst who is familiar with data processing projects and who

wants to use Talend to get your first big data job executed in a reliable, quick, and graphical way, then Talend for Big Data is perfect for you.

Solve business challenges with Microsoft Power BI's advanced visualization and data analysis techniques Key Features Create effective storytelling reports by implementing simple-to-intermediate Power BI features Develop powerful analytical models to extract key insights for changing business needs Build, publish, and share impressive dashboards for your organization Book Description To succeed in today's transforming business world, organizations need business intelligence capabilities to make smarter decisions faster than ever before. This Power BI book is an entry-level guide that will get you up and running with data modeling, visualization, and analytical techniques from scratch. You'll find this book handy if you want to get well-versed with the extensive Power BI ecosystem. You'll start by covering the basics of business intelligence and installing Power BI. You'll then learn the wide range of Power BI features to unlock business insights. As you progress, the book will take you through how to use Power Query to ingest, cleanse, and shape your data, and use Power BI DAX to create simple to complex calculations. You'll also be able to add a variety of interactive visualizations to your reports to bring your data to life. Finally, you'll gain hands-on experience in creating visually stunning reports that speak to business decision makers, and see how you can securely share these reports and collaborate with others. By the end of this book, you'll be ready to create simple, yet effective, BI reports and dashboards using the latest features of Power BI. What you will learn Explore the different features of Power BI to create interactive dashboards Use the Query Editor to import and transform data Perform simple and complex DAX calculations to enhance analysis Discover business insights and tell a story with your data

using Power BI Explore data and learn to manage datasets, dataflows, and data gateways Use workspaces to collaborate with others and publish your reports Who this book is for If you're an IT manager, data analyst, or BI user new to using Power BI for solving business intelligence problems, this book is for you. You'll also find this book useful if you want to migrate from other BI tools to create powerful and interactive dashboards. No experience of working with Power BI is expected.