

Data Lake Development With Big Data

Many enterprises are investing in a next-generation data lake, hoping to democratize data at scale to provide business insights and ultimately make automated intelligent decisions. In this practical book, author Zhamak Dehghani reveals that, despite the time, money, and effort poured into them, data warehouses and data lakes fail when applied at the scale and speed of today's organizations. A distributed data mesh is a better choice. Dehghani guides architects, technical leaders, and decision makers on their journey from monolithic big data architecture to a paradigm that draws from modern distributed architecture. A data mesh considers domains as a first-class concern, applies platform thinking to create self-serve data infrastructure, and treats data as a product. This book shows you why and how. Examine the current landscape of data architectures, their underlying characteristics, and failure modes Learn how to divide data (and its supporting technology stacks and architecture) into operational data and analytical data Get a complete introduction to data mesh principles and logical architecture Create a foundation for gaining value from analytical data and historical facts at scale Move beyond a monolithic data lake to a distributed

data mesh

Organizations invest incredible amounts of time and money obtaining and then storing big data in data stores called data lakes. But how many of these organizations can actually get the data back out in a useable form? Very few can turn the data lake into an information gold mine. Most wind up with garbage dumps. Data Lake Architecture will explain how to build a useful data lake, where data scientists and data analysts can solve business challenges and identify new business opportunities. Learn how to structure data lakes as well as analog, application, and text-based data ponds to provide maximum business value. Understand the role of the raw data pond and when to use an archival data pond. Leverage the four key ingredients for data lake success: metadata, integration mapping, context, and metaprocess. Bill Inmon opened our eyes to the architecture and benefits of a data warehouse, and now he takes us to the next level of data lake architecture.

Software Architecture for Big Data and the Cloud is designed to be a single resource that brings together research on how software architectures can solve the challenges imposed by building big data software systems. The challenges of big data on the software architecture can relate to scale, security, integrity, performance, concurrency, parallelism, and dependability, amongst others. Big data handling

requires rethinking architectural solutions to meet functional and non-functional requirements related to volume, variety and velocity. The book's editors have varied and complementary backgrounds in requirements and architecture, specifically in software architectures for cloud and big data, as well as expertise in software engineering for cloud and big data. This book brings together work across different disciplines in software engineering, including work expanded from conference tracks and workshops led by the editors. Discusses systematic and disciplined approaches to building software architectures for cloud and big data with state-of-the-art methods and techniques Presents case studies involving enterprise, business, and government service deployment of big data applications Shares guidance on theory, frameworks, methodologies, and architecture for cloud and big data

Together, big data and analytics have tremendous potential to improve the way we use precious resources, to provide more personalized services, and to protect ourselves from unexpected and ill-intentioned activities. To fully use big data and analytics, an organization needs a system of insight. This is an ecosystem where individuals can locate and access data, and build visualizations and new analytical models that can be deployed into the IT systems to improve the operations of the organization. The data that is most valuable for

analytics is also valuable in its own right and typically contains personal and private information about key people in the organization such as customers, employees, and suppliers. Although universal access to data is desirable, safeguards are necessary to protect people's privacy, prevent data leakage, and detect suspicious activity. The data reservoir is a reference architecture that balances the desire for easy access to data with information governance and security. The data reservoir reference architecture describes the technical capabilities necessary for a system of insight, while being independent of specific technologies. Being technology independent is important, because most organizations already have investments in data platforms that they want to incorporate in their solution. In addition, technology is continually improving, and the choice of technology is often dictated by the volume, variety, and velocity of the data being managed. A system of insight needs more than technology to succeed. The data reservoir reference architecture includes description of governance and management processes and definitions to ensure the human and business systems around the technology support a collaborative, self-service, and safe environment for data use. The data reservoir reference architecture was first introduced in *Governing and Managing Big Data for Analytics and Decision Makers*,

REDP-5120, which is available at: <http://www.redbooks.ibm.com/redpieces/abstracts/redp5120.html>. This IBM® Redbooks publication, *Designing and Operating a Data Reservoir*, builds on that material to provide more detail on the capabilities and internal workings of a data reservoir.

A handy reference guide for data analysts and data scientists to help to obtain value from big data analytics using Spark on Hadoop clusters

About This Book This book is based on the latest 2.0 version of Apache Spark and 2.7 version of Hadoop integrated with most commonly used tools. Learn all Spark stack components including latest topics such as DataFrames, DataSets, GraphFrames, Structured Streaming, DataFrame based ML Pipelines and SparkR. Integrations with frameworks such as HDFS, YARN and tools such as Jupyter, Zeppelin, NiFi, Mahout, HBase Spark Connector, GraphFrames, H2O and Hivemall.

Who This Book Is For Though this book is primarily aimed at data analysts and data scientists, it will also help architects, programmers, and practitioners. Knowledge of either Spark or Hadoop would be beneficial. It is assumed that you have basic programming background in Scala, Python, SQL, or R programming with basic Linux experience. Working experience within big data environments is not mandatory.

What You Will Learn Find out and implement the tools and techniques of big data

Get Free Data Lake Development With Big Data

analytics using Spark on Hadoop clusters with wide variety of tools used with Spark and Hadoop Understand all the Hadoop and Spark ecosystem components Get to know all the Spark components: Spark Core, Spark SQL, DataFrames, DataSets, Conventional and Structured Streaming, MLLib, ML Pipelines and Graphx See batch and real-time data analytics using Spark Core, Spark SQL, and Conventional and Structured Streaming Get to grips with data science and machine learning using MLLib, ML Pipelines, H2O, Hivemall, Graphx, SparkR and Hivemall. In Detail Big Data Analytics book aims at providing the fundamentals of Apache Spark and Hadoop. All Spark components – Spark Core, Spark SQL, DataFrames, Data sets, Conventional Streaming, Structured Streaming, MLLib, Graphx and Hadoop core components – HDFS, MapReduce and Yarn are explored in greater depth with implementation examples on Spark + Hadoop clusters. It is moving away from MapReduce to Spark. So, advantages of Spark over MapReduce are explained at great depth to reap benefits of in-memory speeds. DataFrames API, Data Sources API and new Data set API are explained for building Big Data analytical applications. Real-time data analytics using Spark Streaming with Apache Kafka and HBase is covered to help building streaming applications. New Structured streaming concept is explained with an IOT (Internet of Things) use case.

Get Free Data Lake Development With Big Data

Machine learning techniques are covered using MLLib, ML Pipelines and SparkR and Graph Analytics are covered with GraphX and GraphFrames components of Spark. Readers will also get an opportunity to get started with web based notebooks such as Jupyter, Apache Zeppelin and data flow tool Apache NiFi to analyze and visualize data. Style and approach This step-by-step pragmatic guide will make life easy no matter what your level of experience. You will deep dive into Apache Spark on Hadoop clusters through ample exciting real-life examples. Practical tutorial explains data science in simple terms to help programmers and data analysts get started with Data Science Data Lake Development with Big DataPackt Publishing Ltd

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

Get Free Data Lake Development With Big Data

environment that includes data security, high availability, data processing, data streaming, and more. Each chapter includes application of a concept, code snippets, and use case demonstrations to provide you with a practical approach. You will learn the concept, scope, application, and starting point. What You'll Learn Get to know data lake architecture and design principles Implement data capture and streaming strategies Implement data processing strategies in Hadoop Understand the data lake security framework and availability model Who This Book Is For Big data architects and solution architects

Helps users understand the breadth of Azure services by organizing them into a reference framework they can use when crafting their own big-data analytics solution.

Explore architectural approaches to building Data Lakes that ingest, index, manage, and analyze massive amounts of data using Big Data technologies About This Book Comprehend the intricacies of architecting a Data Lake and build a data strategy around your current data architecture Efficiently manage vast amounts of data and deliver it to multiple applications and systems with a high degree of performance and scalability Packed with industry best practices and use-case scenarios to get you up-and-running Who This Book Is For This book is for architects and senior managers who are responsible for building a strategy around their current data architecture, helping them identify the need for a Data Lake implementation in an enterprise context. The reader will need a good knowledge of master data management and

Get Free Data Lake Development With Big Data

information lifecycle management, and experience of Big Data technologies. What You Will Learn Identify the need for a Data Lake in your enterprise context and learn to architect a Data Lake Learn to build various tiers of a Data Lake, such as data intake, management, consumption, and governance, with a focus on practical implementation scenarios Find out the key considerations to be taken into account while building each tier of the Data Lake Understand Hadoop-oriented data transfer mechanism to ingest data in batch, micro-batch, and real-time modes Explore various data integration needs and learn how to perform data enrichment and data transformations using Big Data technologies Enable data discovery on the Data Lake to allow users to discover the data Discover how data is packaged and provisioned for consumption Comprehend the importance of including data governance disciplines while building a Data Lake In Detail A Data Lake is a highly scalable platform for storing huge volumes of multistructured data from disparate sources with centralized data management services. This book explores the potential of Data Lakes and explores architectural approaches to building data lakes that ingest, index, manage, and analyze massive amounts of data using batch and real-time processing frameworks. It guides you on how to go about building a Data Lake that is managed by Hadoop and accessed as required by other Big Data applications. This book will guide readers (using best practices) in developing Data Lake's capabilities. It will focus on architect data governance, security, data quality, data lineage tracking, metadata management, and semantic data tagging. By the end of this book, you will have a good understanding of building a Data Lake for Big Data. Style and approach Data Lake Development with Big Data provides architectural approaches to building a Data Lake. It follows a use case-based approach where practical implementation scenarios of

Get Free Data Lake Development With Big Data

each key component are explained. It also helps you understand how these use cases are implemented in a Data Lake. The chapters are organized in a way that mimics the sequential data flow evidenced in a Data Lake.

Aimed at helping business and IT managers clearly communicate with each other, this helpful book addresses concerns straight-on and provides practical methods to building a collaborative data warehouse . You'll get clear explanations of the goals and objectives of each stage of the data warehouse lifecycle while learning the roles that both business managers and technicians play at each stage. Discussions of the most critical decision points for success at each phase of the data warehouse lifecycle help you understand ways in which both business and IT management can make decisions that best meet unified objectives.

AI may be the greatest opportunity of our time, with the potential to add nearly \$16 trillion to the global economy over the next decade. But so far, adoption has been much slower than anticipated, or so headlines may lead you to believe. With this practical guide, business leaders will discover where they are in their AI journey and learn the steps necessary to successfully scale AI throughout their organization. Authors Rob Thomas and Paul Zikopoulos from IBM introduce C-suite executives and business professionals to the AI Ladder—a unified, prescriptive approach to help them understand and accelerate the AI journey. Complete with real-world examples and real-life experiences, this book explores AI drivers, value, and opportunity, as well as the adoption challenges organizations face. Understand why you can't have AI without an information architecture (IA) Appreciate how AI is as much a cultural change as it is a technological one Collect data and make it simple and accessible, regardless of where it lives Organize data to create a business-ready analytics foundation Analyze data, and build and scale AI with trust and

Get Free Data Lake Development With Big Data

transparency Infuse AI throughout your entire business and create intelligent workflows

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

This book constitutes the thoroughly refereed post-conference proceedings of the Satellite Events of the 16th Extended Semantic Web Conference, ESWC 2019, held in Portorož, Slovenia, in June 2019. The volume contains 38 poster and demonstration papers, 2 workshop papers, 5 PhD symposium papers, and 3 industry track papers, selected out of a total of 68 submissions. They deal with all areas of semantic web research, semantic technologies on the Web and Linked Data.

Integrate big data into business to drive competitive advantage and sustainable success Big Data MBA brings

Get Free Data Lake Development With Big Data

insight and expertise to leveraging big data in business so you can harness the power of analytics and gain a true business advantage. Based on a practical framework with supporting methodology and hands-on exercises, this book helps identify where and how big data can help you transform your business. You'll learn how to exploit new sources of customer, product, and operational data, coupled with advanced analytics and data science, to optimize key processes, uncover monetization opportunities, and create new sources of competitive differentiation. The discussion includes guidelines for operationalizing analytics, optimal organizational structure, and using analytic insights throughout your organization's user experience to customers and front-end employees alike. You'll learn to “think like a data scientist” as you build upon the decisions your business is trying to make, the hypotheses you need to test, and the predictions you need to produce. Business stakeholders no longer need to relinquish control of data and analytics to IT. In fact, they must champion the organization's data collection and analysis efforts. This book is a primer on the business approach to analytics, providing the practical understanding you need to convert data into opportunity. Understand where and how to leverage big data Integrate analytics into everyday operations Structure your organization to drive analytic insights Optimize processes, uncover opportunities, and stand out from the rest Help business stakeholders to “think like a data scientist” Understand appropriate business application of different analytic techniques If you want data to transform your business, you need to know how to put it to use. Big Data MBA shows you how to implement big data and analytics to make better decisions.

While many companies ponder implementation details such as distributed processing engines and algorithms for data analysis, this practical book takes a much wider view of big

Get Free Data Lake Development With Big Data

data development, starting with initial planning and moving diligently toward execution. Authors Ted Malaska and Jonathan Seidman guide you through the major components necessary to start, architect, and develop successful big data projects. Everyone from CIOs and COOs to lead architects and developers will explore a variety of big data architectures and applications, from massive data pipelines to web-scale applications. Each chapter addresses a piece of the software development life cycle and identifies patterns to maximize long-term success throughout the life of your project. Start the planning process by considering the key data project types Use guidelines to evaluate and select data management solutions Reduce risk related to technology, your team, and vague requirements Explore system interface design using APIs, REST, and pub/sub systems Choose the right distributed storage system for your big data system Plan and implement metadata collections for your data architecture Use data pipelines to ensure data integrity from source to final storage Evaluate the attributes of various engines for processing the data you collect

The need to handle increasingly larger data volumes is one factor driving the adoption of a new class of nonrelational “NoSQL” databases. Advocates of NoSQL databases claim they can be used to build systems that are more performant, scale better, and are easier to program. NoSQL Distilled is a concise but thorough introduction to this rapidly emerging technology. Pramod J. Sadalage and Martin Fowler explain how NoSQL databases work and the ways that they may be a superior alternative to a traditional RDBMS. The authors provide a fast-paced guide to the concepts you need to know in order to evaluate whether NoSQL databases are right for your needs and, if so, which technologies you should explore further. The first part of the book concentrates on core concepts, including schemaless data models, aggregates,

Get Free Data Lake Development With Big Data

new distribution models, the CAP theorem, and map-reduce. In the second part, the authors explore architectural and design issues associated with implementing NoSQL. They also present realistic use cases that demonstrate NoSQL databases at work and feature representative examples using Riak, MongoDB, Cassandra, and Neo4j. In addition, by drawing on Pramod Sadalage's pioneering work, NoSQL Distilled shows how to implement evolutionary design with schema migration: an essential technique for applying NoSQL databases. The book concludes by describing how NoSQL is ushering in a new age of Polyglot Persistence, where multiple data-storage worlds coexist, and architects can choose the technology best optimized for each type of data access. By 2020, experts forecast that up to 28 billion devices will be connected to the Internet with only one third of them being computers, smartphones and tablets. The remaining two thirds will be other "devices"--sensors, terminals, household appliances, thermostats, televisions, automobiles, production machinery, urban infrastructure and many other "things"--which traditionally have not been Internet enabled. This "Internet of Things" (IoT) represents a remarkable transformation of the way in which our world will soon interact. Much like the World Wide Web connected computers to networks, and the next evolution connected people to the Internet and other people, IoT looks poised to interconnect devices, people, environments, virtual objects and machines in ways that only science fiction writers could have imagined. In a nutshell, the Internet of Things (IoT) is the convergence of connecting people, things, data

and processes. It is transforming our life, business and everything in between. Secure and Smart Internet of Things explores many aspects of the Internet of Things and explains many of the completed principles of IoT and the new advances in IoT including the use of Fog Computing, AI, and Blockchain technology. The topics discussed in the book include: - Internet of Things (IoT) - Industrial Internet of Things (IIoT) - Fog Computing - Artificial Intelligence - Blockchain Technology - Network Security - Zero-Trust Model - Data Analytics - Digital Transformation - DDoS - Smart Devices

"This book discusses various aspects of Industry 4.0 from the perspective of information system evolution. Industry 4.0 refers to a new phase in the industrial revolution that relies heavily on interconnectivity, automation, machine learning, real-time data, the Internet of Things and blockchain technology. The interdisciplinary book addresses a number of topics related to modern information technologies, and presents innovative concepts, methods, models and tools for the development of information systems to support Industry 4.0. Focusing on artificial intelligence, collective knowledge processing and blockchain technology, it appeals to a wide readership, including researchers, students, business managers and professionals, software developers, as well as IT and management specialists. ." -- Prové de l'editor.

Get Free Data Lake Development With Big Data

Learn how to build a data science technology stack and perform good data science with repeatable methods. You will learn how to turn data lakes into business assets. The data science technology stack demonstrated in Practical Data Science is built from components in general use in the industry. Data scientist Andreas Vermeulen demonstrates in detail how to build and provision a technology stack to yield repeatable results. He shows you how to apply practical methods to extract actionable business knowledge from data lakes consisting of data from a polyglot of data types and dimensions. What You'll Learn Become fluent in the essential concepts and terminology of data science and data engineering Build and use a technology stack that meets industry criteria Master the methods for retrieving actionable business knowledge Coordinate the handling of polyglot data types in a data lake for repeatable results Who This Book Is For Data scientists and data engineers who are required to convert data from a data lake into actionable knowledge for their business, and students who aspire to be data scientists and data engineers

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

Get Free Data Lake Development With Big Data

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

The big data era is upon us: data are being generated, analyzed, and used at an unprecedented scale, and data-driven decision making is sweeping through all aspects of society. Since the value of data explodes when it can be linked and fused with other data, addressing the big data integration (BDI) challenge is critical to realizing the promise of big data. BDI differs from traditional data integration along the dimensions of volume, velocity, variety,

and veracity. First, not only can data sources contain a huge volume of data, but also the number of data sources is now in the millions. Second, because of the rate at which newly collected data are made available, many of the data sources are very dynamic, and the number of data sources is also rapidly exploding. Third, data sources are extremely heterogeneous in their structure and content, exhibiting considerable variety even for substantially similar entities. Fourth, the data sources are of widely differing qualities, with significant differences in the coverage, accuracy and timeliness of data provided. This book explores the progress that has been made by the data integration community on the topics of schema alignment, record linkage and data fusion in addressing these novel challenges faced by big data integration. Each of these topics is covered in a systematic way: first starting with a quick tour of the topic in the context of traditional data integration, followed by a detailed, example-driven exposition of recent innovative techniques that have been proposed to address the BDI challenges of volume, velocity, variety, and veracity. Finally, it presents merging topics and opportunities that are specific to BDI, identifying promising directions for the data integration community.

The volume LNCS 12393 constitutes the papers of the 22nd International Conference Big Data Analytics and Knowledge Discovery which will be

held online in September 2020. The 15 full papers presented together with 14 short papers plus 1 position paper in this volume were carefully reviewed and selected from a total of 77 submissions. This volume offers a wide range to following subjects on theoretical and practical aspects of big data analytics and knowledge discovery as a new generation of big data repository, data pre-processing, data mining, text mining, sequences, graph mining, and parallel processing.

The data lakehouse is the next generation of the data warehouse and data lake, designed to meet today's complex and ever-changing analytics, machine learning, and data science requirements. Learn about the features and architecture of the data lakehouse, along with its powerful analytical infrastructure. Appreciate how the universal common connector blends structured, textual, analog, and IoT data. Maintain the lakehouse for future generations through Data Lakehouse Housekeeping and Data Future-proofing. Know how to incorporate the lakehouse into an existing data governance strategy. Incorporate data catalogs, data lineage tools, and open source software into your architecture to ensure your data scientists, analysts, and end users live happily ever after.

Enterprises are experimenting with using Hadoop to build Big Data Lakes, but many projects are stalling or failing because the approaches that worked at

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

Find the right big data solution for your business or organization Big data management is one of the major challenges facing business, industry, and not-for-profit organizations. Data sets such as customer transactions for a mega-retailer, weather patterns monitored by meteorologists, or social network activity can quickly outpace the capacity of traditional data management tools. If you need to develop or manage big data solutions, you'll appreciate how these four experts define, explain, and guide you through this new and often confusing concept. You'll learn what it is, why it matters, and

how to choose and implement solutions that work. Effectively managing big data is an issue of growing importance to businesses, not-for-profit organizations, government, and IT professionals. Authors are experts in information management, big data, and a variety of solutions. Explains big data in detail and discusses how to select and implement a solution, security concerns to consider, data storage and presentation issues, analytics, and much more. Provides essential information in a no-nonsense, easy-to-understand style that is empowering. *Big Data For Dummies* cuts through the confusion and helps you take charge of big data solutions for your organization.

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

Data Warehousing in the Age of the Big Data will help you and your organization make the most of unstructured data with your existing data warehouse. As Big Data continues to revolutionize how we use data, it doesn't have to create more confusion. Expert author Krish Krishnan helps you make sense of how Big Data fits into the world of data warehousing in clear and concise detail. The book is presented in three distinct parts. Part 1 discusses Big Data, its technologies and use cases from early adopters. Part 2 addresses data warehousing, its shortcomings, and new architecture options, workloads, and integration techniques for Big Data and the data warehouse. Part 3 deals with data governance, data visualization, information life-cycle management, data scientists, and implementing a Big Data–ready data warehouse. Extensive appendixes include case studies from vendor implementations and a special segment on how we can build a healthcare information factory. Ultimately, this book will help you navigate through the complex layers of Big Data and data warehousing while providing you information on how

to effectively think about using all these technologies and the architectures to design the next-generation data warehouse. Learn how to leverage Big Data by effectively integrating it into your data warehouse. Includes real-world examples and use cases that clearly demonstrate Hadoop, NoSQL, HBASE, Hive, and other Big Data technologies Understand how to optimize and tune your current data warehouse infrastructure and integrate newer infrastructure matching data processing workloads and requirements

Data-driven insights are a key competitive advantage for any industry today, but deriving insights from raw data can still take days or weeks. Most organizations can't scale data science teams fast enough to keep up with the growing amounts of data to transform. What's the answer? Self-service data. With this practical book, data engineers, data scientists, and team managers will learn how to build a self-service data science platform that helps anyone in your organization extract insights from data. Sandeep Uttamchandani provides a scorecard to track and address bottlenecks that slow down time to insight across data discovery, transformation, processing, and production. This book bridges the gap between data scientists bottlenecked by engineering realities and data engineers unclear about ways to make self-service work. Build a self-service portal to support data discovery, quality,

lineage, and governance Select the best approach for each self-service capability using open source cloud technologies Tailor self-service for the people, processes, and technology maturity of your data platform Implement capabilities to democratize data and reduce time to insight Scale your self-service portal to support a large number of users within your organization

A concise introduction to the emerging field of data science, explaining its evolution, relation to machine learning, current uses, data infrastructure issues, and ethical challenges. The goal of data science is to improve decision making through the analysis of data. Today data science determines the ads we see online, the books and movies that are recommended to us online, which emails are filtered into our spam folders, and even how much we pay for health insurance. This volume in the MIT Press Essential Knowledge series offers a concise introduction to the emerging field of data science, explaining its evolution, current uses, data infrastructure issues, and ethical challenges. It has never been easier for organizations to gather, store, and process data. Use of data science is driven by the rise of big data and social media, the development of high-performance computing, and the emergence of such powerful methods for data analysis and modeling as deep learning. Data science encompasses a set of principles, problem definitions, algorithms, and

processes for extracting non-obvious and useful patterns from large datasets. It is closely related to the fields of data mining and machine learning, but broader in scope. This book offers a brief history of the field, introduces fundamental data concepts, and describes the stages in a data science project. It considers data infrastructure and the challenges posed by integrating data from multiple sources, introduces the basics of machine learning, and discusses how to link machine learning expertise with real-world problems. The book also reviews ethical and legal issues, developments in data regulation, and computational approaches to preserving privacy. Finally, it considers the future impact of data science and offers principles for success in data science projects.

Data mining of massive data sets is transforming the way we think about crisis response, marketing, entertainment, cybersecurity and national intelligence. Collections of documents, images, videos, and networks are being thought of not merely as bit strings to be stored, indexed, and retrieved, but as potential sources of discovery and knowledge, requiring sophisticated analysis techniques that go far beyond classical indexing and keyword counting, aiming to find relational and semantic interpretations of the phenomena underlying the data. *Frontiers in Massive Data Analysis* examines the frontier of analyzing massive

amounts of data, whether in a static database or streaming through a system. Data at that scale--terabytes and petabytes--is increasingly common in science (e.g., particle physics, remote sensing, genomics), Internet commerce, business analytics, national security, communications, and elsewhere. The tools that work to infer knowledge from data at smaller scales do not necessarily work, or work well, at such massive scale. New tools, skills, and approaches are necessary, and this report identifies many of them, plus promising research directions to explore. *Frontiers in Massive Data Analysis* discusses pitfalls in trying to infer knowledge from massive data, and it characterizes seven major classes of computation that are common in the analysis of massive data. Overall, this report illustrates the cross-disciplinary knowledge--from computer science, statistics, machine learning, and application disciplines--that must be brought to bear to make useful inferences from massive data.

Get a 360-degree view of how the journey of data analytics solutions has evolved from monolithic data stores and enterprise data warehouses to data lakes and modern data warehouses. You will This book includes comprehensive coverage of how: To architect data lake analytics solutions by choosing suitable technologies available on Microsoft Azure The advent of microservices applications covering

ecommerce or modern solutions built on IoT and how real-time streaming data has completely disrupted this ecosystem. These data analytics solutions have been transformed from solely understanding the trends from historical data to building predictions by infusing machine learning technologies into the solutions. Data platform professionals who have been working on relational data stores, non-relational data stores, and big data technologies will find the content in this book useful. The book also can help you start your journey into the data engineer world as it provides an overview of advanced data analytics and touches on data science concepts and various artificial intelligence and machine learning technologies available on Microsoft Azure.

What Will You Learn

You will understand the: Concepts of data lake analytics, the modern data warehouse, and advanced data analytics Architecture patterns of the modern data warehouse and advanced data analytics solutions Phases—such as Data Ingestion, Store, Prep and Train, and Model and Serve—of data analytics solutions and technology choices available on Azure under each phase In-depth coverage of real-time and batch mode data analytics solutions architecture Various managed services available on Azure such as Synapse analytics, event hubs, Stream analytics, CosmosDB, and managed Hadoop services such as Databricks and HDInsight Who This Book Is For

Data platform professionals, database architects, engineers, and solution architects

This book contains practical steps business users can take to implement data management in a number of ways, including data governance, data architecture, master data management, business intelligence, and others. It defines data strategy, and covers chapters that illustrate how to align a data strategy with the business strategy, a discussion on valuing data as an asset, the evolution of data management, and who should oversee a data strategy. This provides the user with a good understanding of what a data strategy is and its limits. Critical to a data strategy is the incorporation of one or more data management domains.

Chapters on key data management domains—data governance, data architecture, master data management and analytics, offer the user a practical approach to data management execution within a data strategy. The intent is to enable the user to identify how execution on one or more data management domains can help solve business issues. This book is intended for business users who work with data, who need to manage one or more aspects of the organization's data, and who want to foster an integrated approach for how enterprise data is managed. This book is also an excellent reference for students studying computer science and business management or simply for someone

who has been tasked with starting or improving existing data management.

This IBM Redguide™ publication looks back on the key decisions that made the data lake successful and looks forward to the future. It proposes that the metadata management and governance approaches developed for the data lake can be adopted more broadly to increase the value that an organization gets from its data. Delivering this broader vision, however, requires a new generation of data catalogs and governance tools built on open standards that are adopted by a multi-vendor ecosystem of data platforms and tools. Work is already underway to define and deliver this capability, and there are multiple ways to engage. This guide covers the reasons why this new capability is critical for modern businesses and how you can get value from it.

This book constitutes the proceedings of the 8th International Conference on Big Data Analytics, BDA 2020, which took place during December 15-18, 2020, in Sonapat, India. The 11 full and 3 short papers included in this volume were carefully reviewed and selected from 48 submissions; the book also contains 4 invited and 3 tutorial papers. The contributions were organized in topical sections named as follows: data science systems; data science architectures; big data analytics in healthcare; information interchange of Web data resources; and business analytics.

With this practical book, AI and machine learning practitioners will learn how to successfully build and deploy data science projects on Amazon Web Services. The Amazon AI and machine learning stack unifies data science, data engineering, and application development to help level up your skills. This guide shows you how to build and run pipelines in the cloud, then integrate the results into applications in minutes instead of days. Throughout the book,

Get Free Data Lake Development With Big Data

authors Chris Fregly and Antje Barth demonstrate how to reduce cost and improve performance. Apply the Amazon AI and ML stack to real-world use cases for natural language processing, computer vision, fraud detection, conversational devices, and more Use automated machine learning to implement a specific subset of use cases with SageMaker Autopilot Dive deep into the complete model development lifecycle for a BERT-based NLP use case including data ingestion, analysis, model training, and deployment Tie everything together into a repeatable machine learning operations pipeline Explore real-time ML, anomaly detection, and streaming analytics on data streams with Amazon Kinesis and Managed Streaming for Apache Kafka Learn security best practices for data science projects and workflows including identity and access management, authentication, authorization, and more

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

Get Free Data Lake Development With Big Data

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

Get Free Data Lake Development With Big Data

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.

Get ready to unlock the power of your data. With the fourth edition of this comprehensive guide, you'll learn how to build and maintain reliable, scalable, distributed systems with Apache Hadoop. This book is ideal for programmers looking to analyze datasets of any size, and for administrators who want to set up and run Hadoop clusters. Using Hadoop 2 exclusively, author Tom White presents new chapters on YARN and several Hadoop-related projects such as Parquet, Flume, Crunch, and Spark. You'll learn about recent changes to Hadoop, and explore new case studies on Hadoop's role in healthcare systems and genomics data processing. Learn fundamental components such as MapReduce, HDFS, and YARN Explore MapReduce in depth, including steps for developing applications with it Set up and maintain a Hadoop cluster running HDFS and MapReduce on YARN Learn two data formats: Avro for data serialization and Parquet for nested data Use data ingestion tools such as Flume (for streaming data) and Sqoop (for bulk data transfer) Understand how high-level data processing tools like Pig, Hive, Crunch, and Spark work with Hadoop Learn the HBase distributed database and the ZooKeeper distributed configuration service

This book presents unique techniques to conquer different Big Data processing and analytics challenges using Hadoop. Practical examples are provided to boost your understanding of Big Data concepts and their implementation. By the end of the book, you will have all the knowledge and skills you need to become a true Big Data expert.

As data management and integration continue to evolve rapidly, storing all your data in one place, such as a data warehouse, is no longer **scalable**. In the very near future,

Get Free Data Lake Development With Big Data

data will need to be distributed and available for several technological solutions. With this practical book, you'll learn how to migrate your enterprise from a complex and tightly coupled data landscape to a more flexible architecture ready for the modern world of data consumption. Executives, data architects, analytics teams, and compliance and governance staff will learn how to build a modern scalable data landscape using the Scaled Architecture, which you can introduce incrementally without a large upfront investment. Author Piethen Strengholt provides blueprints, principles, observations, best practices, and patterns to get you up to speed. Examine data management trends, including technological developments, regulatory requirements, and privacy concerns Go deep into the Scaled Architecture and learn how the pieces fit together Explore data governance and data security, master data management, self-service data marketplaces, and the importance of metadata

Due to the growing use of web applications and communication devices, the use of data has increased throughout various industries. It is necessary to develop new techniques for managing data in order to ensure adequate usage. The Handbook of Research on Pattern Engineering System Development for Big Data Analytics is a critical scholarly resource that examines the incorporation of pattern management in business technologies as well as decision making and prediction process through the use of data management and analysis. Featuring coverage on a broad range of topics such as business intelligence, feature extraction, and data collection, this publication is geared towards professionals, academicians, practitioners, and researchers seeking current research on the development of pattern management systems for business applications.

[Copyright: 10bb6ca1e4dd661cf01a02eeb8885484](https://doi.org/10.1002/9781119488548.ch10)