Data Lake Development With Big Data

Explore the concepts of functional programming, data streaming, and machine learning
Data Lake Development with Big Data
Accelerate Your Journey to AI
Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data
ESWC 2019 Satellite Events, Portorož, Slovenia, June 2–6, 2019, Revised Selected Papers
NoSQLand Big Data
Foundations for Architecting Data Solutions
Delivering Data-Driven Value at Scale
Conceptual Modeling Perspectives
A Guide to Building the Technology Stack for Turning Data Lakes into Business Assets
The Enterprise Big Data Lake
Create scalable pipelines that ingest, curate, and aggregate complex data in a timely and secure way
Modern Enterprise Data Pipelines
Handbook of Research on Pattern Engineering System Development for Big Data Analytics
Data Lakes For Dummies
Trino: The Definitive Guide
Rise of the Data Cloud
A guide for data engineers
How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results
Big Data in Practice
Scala and Spark for Big Data Analytics
Driving Business Strategies with Data Science
Data Lake Architecture
Data Lake Analytics on Microsoft Azure
NoSQL Distilled
Practical Data Science
Azure Storage, Streaming, and Batch Analytics
Managing Successful Data Projects
Building the Data Lakehouse
Big Data Applications in Industry 4. 0
Big Data Integration
A Practitioner's Guide to Big Data Engineering
Knowledge Graphs and Big Data Processing
Data Engineering with Apache Spark, Delta Lake, and Lakehouse
A Brief Guide to the Emerging World of Polyglot Persistence
Towards Industry 4.0 — Current Challenges in Information Systems
Enterprise Big Data Warehouse, BI Implementations and Analytics
Software Architecture for Big Data and the Cloud

Data Mesh
Designing the Data Lake and Avoiding the Garbage Dump

FINLEY CHASE

Explore the concepts of functional programming, data streaming, and machine learning "O'Reilly Media, Inc."
Perform fast interactive analytics against different data sources using the Trino high-performance distributed SQL query engine. With this practical guide, you'll learn how to conduct analytics on data where it lives, whether it's Hive, Cassandra, a relational database, or a proprietary data store. Analysts, software engineers, and production engineers will learn how to manage, use, and even develop with Trino. Initially developed by Facebook, open source Trino is now used by Netflix, Airbnb, LinkedIn, Twitter, Uber, and many other companies. Matt Fuller, Manfred Moser, and Martin Traverso show you how a single Trino query can combine data from multiple sources to allow for analytics across your entire organization. Get started: Explore Trino's use cases and learn about tools that will help you connect to Trino and query data Go deeper: Learn Trino's internal workings, including how to connect to and query data sources with support for SQL statements, operators, functions, and more Put Trino in production: Secure Trino, monitor workloads, tune queries, and connect more applications; learn how other organizations apply Trino
Data Lake Development with Big Data Springer Nature
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
Accelerate Your Journey to AI "O'Reilly Media, Inc."
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, 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.
Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data Springer Nature
Enter the fast-paced world of SAP HANA 2.0 with this introductory guide. Begin with an exploration of the technological backbone of SAP HANA as a database and platform. Then, step into key SAP HANA user roles and discover core capabilities for administration, application development, advanced analytics, security, data integration, and more. No matter how SAP HANA 2.0 fits into your business, this book is your starting point. In this book, you'll learn about: a. Technology Discover what makes an in-memory database platform. Learn about SAP HANA's journey from version 1.0 to 2.0, take a tour of your technology options, and walk through deployment scenarios and implementation requirements. b. Tools Unpack your SAP HANA toolkit. See essential tools in action, from SAP HANA cockpit and SAP HANA studio, to the SAP HANA Predictive Analytics Library and SAP HANA smart data integration. c. Key Roles Understand how to use SAP HANA as a developer, administrator, data scientist, data center architect, and more. Explore key tasks like backend programming with SQLScript, security setup with roles and authorizations, data integration with the SAP HANA Data Management Suite, and more. Highlights include: 1) Architecture 2) Administration 3) Application development 4) Analytics 5) Security 6) Data integration 7) Data architecture 8) Data center
ESWC 2019 Satellite Events, Portorož, Slovenia, June 2–6, 2019, Revised Selected Papers Packt Publishing Ltd
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

NoSQLand Big Data IBM Redbooks

Harness the power of Scala to program Spark and analyze tonnes of data in the blink of an eye! About This Book Learn Scala's sophisticated type system that combines Functional Programming and object-oriented concepts Work on a wide array of applications, from simple batch jobs to stream processing and machine learning Explore the most common as well as some complex use-cases to perform large-scale data analysis with Spark Who This Book Is For Anyone who wishes to learn how to perform data analysis by harnessing the power of Spark will find this book extremely useful. No knowledge of Spark or Scala is assumed, although prior programming experience (especially with other JVM languages) will be useful to pick up concepts quicker. What You Will Learn Understand object-oriented & functional programming concepts of Scala In-depth understanding of Scala collection APIs Work with RDD and DataFrame to learn Spark's core abstractions Analysing structured and unstructured data using SparkSQL and GraphX Scalable and fault-tolerant streaming application development using Spark structured streaming Learn machine-learning best practices for classification, regression, dimensionality reduction, and recommendation system to build predictive models with widely used algorithms in Spark MLlib & ML Build clustering models to cluster a vast amount of data Understand tuning, debugging, and monitoring Spark applications Deploy Spark applications on real clusters in Standalone, Mesos, and YARN In Detail Scala has been observing wide adoption over the past few years, especially in the field of data science and analytics. Spark, built on Scala, has gained a lot of recognition and is being used widely in productions. Thus, if you want to leverage the power of Scala and Spark to make sense of big data, this book is for you. The first part introduces you to Scala, helping you understand the object-oriented and functional programming concepts needed for Spark application development. It then moves on to Spark to cover the basic abstractions using RDD and DataFrame. This will help you develop scalable and fault-tolerant streaming applications by analyzing structured and unstructured data using SparkSQL, GraphX, and Spark structured streaming. Finally, the book moves on to some advanced topics, such as monitoring, configuration, debugging, testing, and deployment. You will also learn how to develop Spark applications using SparkR and PySpark APIs, interactive data analytics using Zeppelin, and in-memory data processing with Alluxio. By the end of this book, you will have a thorough understanding of Spark, and you will be able to perform full-stack data analytics with a feel that no amount of data is too big. Style and approach Filled with practical examples and use cases, this book will hot only help you get up and running with Spark, but will also take you farther down the road to becoming a data scientist.

Foundations for Architecting Data Solutions AuthorHouse

This open access book is part of the LAMBDA Project (Learning, Applying, Multiplying Big Data Analytics), funded by the European Union, GA No. 809965. Data Analytics involves applying algorithmic processes to derive insights. Nowadays it is used in many industries to allow organizations and companies to make better decisions as well as to verify or disprove existing theories or models. The term data analytics is often used interchangeably with intelligence, statistics, reasoning, data mining, knowledge discovery, and others. The goal of this book is to introduce some of the definitions, methods, tools, frameworks, and solutions for big data processing, starting from the process of information extraction and knowledge representation, via knowledge processing and analytics to visualization, sense-making, and practical applications. Each chapter in this book addresses some pertinent aspect of the data processing chain, with a specific focus on understanding Enterprise Knowledge Graphs, Semantic Big Data Architectures, and Smart Data Analytics solutions. This book is addressed to graduate students from technical disciplines, to professional audiences following continuous education short courses, and to researchers from diverse areas following self-study courses. Basic skills in computer science, mathematics, and statistics are required.

Delivering Data-Driven Value at Scale Springer Nature

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.

Conceptual Modeling Perspectives Packt Publishing Ltd

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

A Guide to Building the Technology Stack for Turning Data Lakes into Business Assets Springer

A Dell Technologies perspective on today's data landscape and the key ingredients for planning a modern, distributed data pipeline for your multicloud data-driven enterprise

The Enterprise Big Data Lake "O'Reilly Media, Inc."

The Microsoft Azure cloud is an ideal platform for data-intensive applications. Designed for productivity, Azure provides pre-built services that make collection, storage, and analysis much easier to implement and manage. Azure Storage, Streaming, and Batch Analytics teaches you how to design a reliable, performant, and cost-effective data infrastructure in Azure by progressively building a complete working analytics system. Summary The Microsoft Azure cloud is an ideal platform for data-intensive applications. Designed for productivity, Azure provides pre-built services that make collection, storage, and analysis much easier to implement and manage. Azure Storage, Streaming, and Batch Analytics teaches you how to design a reliable, performant, and cost-effective data infrastructure in Azure by progressively building a complete working analytics system. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Microsoft Azure provides dozens of services that simplify storing and processing data. These services are secure, reliable, scalable, and cost efficient. About the book Azure Storage, Streaming, and Batch Analytics shows you how to build state-of-the-art data solutions with tools from the Microsoft Azure platform. Read along to construct a cloud-native data warehouse, adding features like real-time data processing. Based on the Lambda architecture for big data, the design uses scalable services such as Event Hubs, Stream Analytics, and SQL databases. Along the way, you'll cover most of the topics needed to earn an Azure data engineering certification. What's inside Configuring Azure services for speed and cost Constructing data pipelines with Data Factory Choosing the right data storage methods About the reader For readers familiar with database management. Examples in C# and PowerShell. About the author Richard Nuckolls is a senior developer building big data analytics and reporting systems in Azure. Table of Contents 1 What is data engineering? 2 Building an analytics system in Azure 3 General storage with Azure Storage accounts 4 Azure Data Lake Storage 5 Message handling with Event Hubs 6 Real-time queries with Azure Stream Analytics 7 Batch queries with Azure Data Lake Analytics 8 U-SQL for complex analytics 9 Integrating with Azure Data Lake Analytics 10 Service integration with Azure Data Factory 11 Managed SQL with Azure SQL Database 12 Integrating Data Factory with SQL Database 13 Where to go next

Create scalable pipelines that ingest, curate, and aggregate complex data in a timely and secure way "O'Reilly Media, Inc."

Conceptual modeling has always been one of the main issues in information systems engineering as it aims to describe the general knowledge of the system at an abstract level that facilitates user understanding and software development. This collection of selected papers provides a comprehensive and extremely readable overview of what conceptual modeling is and perspectives on making it more and more relevant in our society. It covers topics like modeling the human genome, blockchain technology, model-driven software development, data integration, and wiki-like repositories and demonstrates the general applicability of conceptual modeling to various problems in diverse domains. Overall, this book is a source of inspiration for everybody in academia working on the vision of creating a strong, fruitful and creative community of conceptual modelers. With this book the editors and authors want to honor Prof. Antoni Olivé for his enormous and ongoing contributions to the conceptual modeling discipline. It was presented to him on the occasion of his keynote at ER 2017 in Valencia, a conference that he has contributed to and supported for over 20 years. Thank you very much to Antoni for so many years of cooperation and friendship.

Modern Enterprise Data Pipelines McGraw Hill Professional

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.

Handbook of Research on Pattern Engineering System Development for Big Data Analytics Pearson Education

"Industry 4.0 is the latest technological innovation in manufacturing with the goal to increase productivity in a flexible and efficient manner. Changing the way in which manufacturers operate, this revolutionary transformation is powered by various technology advances including artificial intelligence (AI), Big Data analytics, Internet-of-Things (IoT) and cloud computing. Big Data analytics has been identified as one of the significant components of Industry 4.0, as it provides valuable insights for smart factory management. Big Data and Industry 4.0 have the potential to reduce resource consumption and optimize processes, thereby playing a key role in achieving sustainable development. Big Data Applications in Industry 4.0 covers the recent advancements that have emerged in the field of Big Data and its applications. The book introduces the concepts and advanced tools and technologies for representing and processing Big Data. It also covers applications of Big Data in such domains as financial services, education, healthcare, biomedical research, logistics, and warehouse management. Researchers, students, scientists, engineers, and statisticians can turn to this book to learn about concepts, technologies, and applications that solve real world problems. The books features: An introduction to data science and the types of data analytics methods accessible today: An overview of data integration concepts, methodologies, and solutions. A general framework of forecasting principles and applications as well as basic forecasting models including naèive, moving average, and exponential smoothing models. A detailed roadmap of the Big Data evolution and its related technological transformation in computing, along with a brief description of related terminologies. The application of Industry 4.0 and Big Data in the field of education. The features, prospects, and significant role of Big Data in banking industry, as well as various use cases of Big Data in banking, finance services, and insurance. Implementing a Data Lake (DL) in the cloud and the significance of a data lake in for decision-making"--

Data Lakes For Dummies Springer Nature

Data Lake Development with Big DataPackt Publishing Ltd

Trino: The Definitive Guide Apress

This IBM RedguideTM 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.

Rise of the Data Cloud CRC Press

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 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

A guide for data engineers Apress

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, data will need to be distributed and available for several technological solutions. With this practical book, you'll learnhow 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 Piethein 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

How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results Manning Publications

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 Sonepat, 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.

Big Data in Practice Apress

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.