Learn Apache Tika Java Technologies

Summary CMIS and Apache Chemistry in Action is a comprehensive guide to the CMIS standard and related ECM concepts, written by the authors of the standard. In it, you'll tackle hands-on examples for building applications on CMIS repositories from both the client and the server sides. You'll learn how to create new content-centric applications that install and run in any CMIS-compliant repository. About The Technology Content Management Interoperability Services (CMIS) is an OASIS standard for accessing content management systems. It specifies a vendor-and language-neutral way to interact with any compliant content repository. Apache Chemistry provides complete reference implementations of the CMIS standard with robust APIs for developers writing tools, applications, and servers. About This Book CMIS and Apache Chemistry in Action is a comprehensive guide to the CMIS standard and related ECM concepts. In it, you'll find clear teaching and instantly useful examples for building content-centric client and server-side applications that run against any CMIS-compliant repository. In fact, using the CMIS Workbench and the InMemory Repository from Apache Chemistry, you'll have running code talking to a real CMIS server by the end of chapter 1. This book requires some familiarity with content management systems and a standard programming language like Java or C#. No exposure to CMIS or Apache Chemistry is assumed. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. What's Inside The only CMIS book endorsed by OASIS Complete coverage of the CMIS 1.0 and 1.1 specifications Cookbook-style tutorials and real-world examples About the Authors Florian Müller, Jay Brown, and Jeff Potts are among the original authors, contributors, and leaders of Apache Chemistry and the OASIS CMIS specification. They continue to shape CMIS implementations at Alfresco, IBM, and SAP. Table of Contents PART 1 UNDERSTANDING CMIS Introducing CMIS Exploring the CMIS domain model Creating, updating, and deleting objects with CMIS CMIS metadata: types and properties Query PART 2 HANDS-ON CMIS CLIENT DEVELOPMENT Meet your new project: The Blend The Blend: read and query functionality The Blend: create, update, and delete functionality Using other client libraries Building mobile apps with CMIS PART 3 ADVANCED TOPICS CMIS bindings Security and control Performance Building a CMIS server

Explore various approaches to organize and extract useful text from unstructured data using Java Key Features Use deep learning and NLP techniques in Java to discover hidden insights in text Work with popular Java libraries such as CoreNLP, OpenNLP, and Mallet Explore machine translation, identifying parts of speech, and topic modeling Book Description Natural Language Processing (NLP) allows you to take any sentence and identify patterns, special names, company names, and more. The second edition of Natural Language Processing with Java teaches you how to perform language analysis with the help of Java libraries, while constantly gaining insights

from the outcomes. You'll start by understanding how NLP and its various concepts work. Having got to grips with the basics, you'll explore important tools and libraries in Java for NLP, such as CoreNLP, OpenNLP, Neuroph, and Mallet. You'll then start performing NLP on different inputs and tasks, such as tokenization, model training, parts-of-speech and parsing trees. You'll learn about statistical machine translation, summarization, dialog systems, complex searches, supervised and unsupervised NLP, and more. By the end of this book, you'll have learned more about NLP, neural networks, and various other trained models in Java for enhancing the performance of NLP applications. What you will learn Understand basic NLP tasks and how they relate to one another Discover and use the available tokenization engines Apply search techniques to find people, as well as things, within a document Construct solutions to identify parts of speech within sentences Use parsers to extract relationships between elements of a document Identify topics in a set of documents Explore topic modeling from a document Who this book is for Natural Language Processing with Java is for you if you are a data analyst, data scientist, or machine learning engineer who wants to extract information from a language using Java. Knowledge of Java programming is needed, while a basic understanding of statistics will be useful but not mandatory.

Ongoing advancements in modern technology have led to significant developments in artificial intelligence. With the numerous applications available, it becomes imperative to conduct research and make further progress in this field. Artificial Intelligence: Concepts, Methodologies, Tools, and Applications provides a comprehensive overview of the latest breakthroughs and recent progress in artificial intelligence. Highlighting relevant technologies, uses, and techniques across various industries and settings, this publication is a pivotal reference source for researchers, professionals, academics, upper-level students, and practitioners interested in emerging perspectives in the field of artificial intelligence.

Learn advanced analytical techniques and leverage existing tool kits to make your analytic applications more powerful, precise, and efficient. This book provides the right combination of architecture, design, and implementation information to create analytical systems that go beyond the basics of classification, clustering, and recommendation. Pro Hadoop Data Analytics emphasizes best practices to ensure coherent, efficient development. A complete example system will be developed using standard third-party components that consist of the tool kits, libraries, visualization and reporting code, as well as support glue to provide a working and extensible end-to-end system. The book also highlights the importance of end-to-end, flexible, configurable, high-performance data pipeline systems with analytical components as well as appropriate visualization results. You'll discover the importance of mix-and-match or hybrid systems, using different analytical components in one application. This hybrid approach will be prominent in the examples. What You'll Learn Build big data analytic systems with the Hadoop ecosystem Use libraries, tool kits, and algorithms to make development easier and more

effective Apply metrics to measure performance and efficiency of components and systems Connect to standard relational databases, noSQL data sources, and more Follow case studies with example components to create your own systems Who This Book Is For Software engineers, architects, and data scientists with an interest in the design and implementation of big data analytical systems using Hadoop, the Hadoop ecosystem, and other associated technologies.

Proceedings of ICMLIP 2020

Learning Spark

Scaling Apache Solr

R in Action, Third Edition

Artificial Intelligence: Concepts, Methodologies, Tools, and Applications

How to Find, Organize, and Manipulate It

Search is everywhere, yet it is one of the most misunderstood functionalities of the IT industry. In Apache Solr, author Xavier Morera guides you through the basics of this highly popular enterprise search tool. You'll learn how to set up an index and how to make it searchable, then query it with a simple enterprise search. Explanations for precision and recall are also included to help you ensure that relevant, accurate results have been returned. Custom UIs using Solritas and SolrNet are also covered. This updated and expanded second edition of Book provides a user-friendly introduction to the subject, Taking a clear structural framework, it guides the reader through the subject's core elements. A flowing writing style combines with the use of illustrations and diagrams throughout the text to ensure the reader understands even the most complex of concepts. This succinct and enlightening overview is a required reading for all those interested in the subject . We hope you find this book useful in shaping your future career & Business. Enhance your Solr indexing experience with advanced techniques and the built-in functionalities available in Apache Solr About This Book Learn about distributed indexing and real-time optimization to change index data on fly Index data from various sources and web crawlers using built-in analyzers and tokenizers This step-by-step guide is packed with real-life examples on indexing data Who This Book Is For This book is for developers who want to increase their experience of indexing in Solr by learning about the various index handlers, analyzers, and methods available in Solr. Beginner level Solr development skills are expected. What You Will Learn Get to know the basic features of Solr indexing and the analyzers/tokenizers available Index XML/JSON data in Solr using the HTTP Post tool and CURL command Work with Data Import Handler to index data from a database Use Apache Tika with Solr to index word documents, PDFs, and much more Utilize Apache Nutch and Solr integration to index crawled data from web pages Update indexes in real-time data feeds Discover techniques to index multi-language and distributed data in Solr Combine the various indexing techniques into a real-life working example of an online shopping web application In Detail Apache Solr is a widely used,

open source enterprise search server that delivers powerful indexing and searching features. These features help fetch relevant information from various sources and documentation. Solr also combines with other open source tools such as Apache Tika and Apache Nutch to provide more powerful features. This fast-paced guide starts by helping you set up Solr and get acquainted with its basic building blocks, to give you a better understanding of Solr indexing. You'll quickly move on to indexing text and boosting the indexing time. Next, you'll focus on basic indexing techniques, various index handlers designed to modify documents, and indexing a structured data source through Data Import Handler. Moving on, you will learn techniques to perform real-time indexing and atomic updates, as well as more advanced indexing techniques such as de-duplication. Later on, we'll help you set up a cluster of Solr servers that combine fault tolerance and high availability. You will also gain insights into working scenarios of different aspects of Solr and how to use Solr with e-commerce data. By the end of the book, you will be competent and confident working with indexing and will have a good knowledge base to efficiently program elements. Style and approach This fast-paced guide is packed with examples that are written in an easy-to-follow style, and are accompanied by detailed explanation. Working examples are included to help you get better results for your applications.

Summary Taming Text, winner of the 2013 Jolt Awards for Productivity, is a hands-on, example-driven guide to working with unstructured text in the context of real-world applications. This book explores how to automatically organize text using approaches such as full-text search, proper name recognition, clustering, tagging, information extraction, and summarization. The book guides you through examples illustrating each of these topics, as well as the foundations upon which they are built. About this Book There is so much text in our lives, we are practically drowningin it. Fortunately, there are innovative tools and techniquesfor managing unstructured information that can throw thesmart developer a much-needed lifeline. You'll find them in thisbook. Taming Text is a practical, example-driven guide to working withtext in real applications. This book introduces you to useful techniques like full-text search, proper name recognition,clustering, tagging, information extraction, and summarization.You'll explore real use cases as you systematically absorb thefoundations upon which they are built.Written in a clear and concise style, this book avoids jargon, explainingthe subject in terms you can understand without a backgroundin statistics or natural language processing. Examples arein Java, but the concepts can be applied in any language. Written for Java developers, the book requires no prior knowledge of GWT. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book. Winner of 2013 Jolt Awards: The Best Books—one of five notable books every serious programmer should read. What's Inside When to use text-taming techniques Important open-source libraries like Solr and Mahout How to build text-processing applications About the Authors Grant Ingersoll is an engineer, speaker, and trainer, a Lucenecommitter, and a cofounder of the Mahout machine-learning project. Thomas Morton is the primary developer of OpenNLP and Maximum Entropy. Drew Farris is a technology consultant, software developer, and contributor to Mahout,Lucene, and Solr. "Takes the mystery out of verycomplex

processes."—From the Foreword by Liz Liddy, Dean, iSchool, Syracuse University Table of Contents Getting started taming text Foundations of taming text Searching Fuzzy string matching Identifying people, places, and things Clustering text Classification, categorization, and tagging Building an example question answering system Untamed text: exploring the next frontier

This book focuses on data and how modern business firms use social data, specifically Online Social Networks (OSNs) incorporated as part of the infrastructure for a number of emerging applications such as personalized recommendation systems, opinion analysis, expertise retrieval, and computational advertising. This book identifies how in such applications, social data offers a plethora of benefits to enhance the decision making process. This book highlights that business intelligence applications are more focused on structured data; however, in order to understand and analyse the social big data, there is a need to aggregate data from various sources and to present it in a plausible format. Big Social Data (BSD) exhibit all the typical properties of big data: wide physical distribution, diversity of formats, non-standard data models, independently-managed and heterogeneous semantics but even further valuable with marketing opportunities. The book provides a review of the current state-of-the-art approaches for big social data analytics as well as to present dissimilar methods to infer value from social data. The book further examines several areas of research that benefits from the propagation of the social data. In particular, the book presents various technical approaches that produce data analytics capable of handling big data features and effective in filtering out unsolicited data and inferring a value. These approaches comprise advanced technical solutions able to capture huge amounts of generated data, scrutinise the collected data to eliminate unwanted data, measure the quality of the inferred data, and transform the amended data for further data analysis. Furthermore, the book presents solutions to derive knowledge and sentiments from BSD and to provide social data classification and prediction. The approaches in this book also incorporate several technologies such as semantic discovery, sentiment analysis, affective computing and machine learning. This book has additional special feature enriched with numerous illustrations such as tables, graphs and charts incorporating advanced visualisation tools in accessible an attractive display.

Digital Transformation of Supply Chain Management

Hadoop: The Definitive Guide

Pro Hadoop Data Analytics

Technological Innovation for the Internet of Things

Introduction to Apache Flink

Designing and Building Big Data Systems using the Hadoop Ecosystem

Deep Learning for Search

This book is a step-by-step guide for readers who would like to learn how to build complete enterprise search solutions, with ample real-world examples and case studies. If you are a developer, designer, or architect who would like to build enterprise search solutions for your

customers or organization, but have no prior knowledge of Apache Solr/Lucene technologies, this is the book for you.

This book constitutes the refereed proceedings of the 4th IFIP WG 5.5/SOCOLNET Doctoral Conference on Computing, Electrical and Industrial Systems, DoCEIS 2013, held in Costa de Caparica, Portugal, in April 2013. The 69 revised full papers were carefully reviewed and selected from numerous submissions. They cover a wide spectrum of topics ranging from collaborative enterprise networks to microelectronics. The papers are organized in the following topical sections: collaborative enterprise networks; service orientation; intelligent computational systems; computational systems; computational systems applications; perceptional systems; robotics and manufacturing; embedded systems and Petri nets; control and decision; integration of power electronics systems with ICT; energy generation; energy distribution; energy transformation; optimization techniques in energy; telecommunications; electronics: devices design; electronics: amplifiers; electronics: RF applications; and electronics: applications.

Accelerate your enterprise search engine and bring relevancy in your search analytics Key Features A practical guide in building expertise with Indexing, Faceting, Clustering and Pagination Master the management and administration of Enterprise Search Applications and services seamlessly Handle multiple data inputs such as JSON, xml, pdf, doc, xls,ppt, csv and much more. Book Description Apache Solr is the only standalone enterprise search server with a REST-like application interface. providing highly scalable, distributed search and index replication for many of the world's largest internet sites. To begin with, you would be introduced to how you perform full text search, multiple filter search, perform dynamic clustering and so on helping you to brush up the basics of Apache Solr. You will also explore the new features and advanced options released in Apache Solr 7.x which will get you numerous performance aspects and making data investigation simpler, easier and powerful. You will learn to build complex queries, extensive filters and how are they compiled in your system to bring relevance in your search tools. You will learn to carry out Solr scoring, elements affecting the document score and how you can optimize or tune the score for the application at hand. You will learn to extract features of documents, writing complex queries in re-ranking the documents. You will also learn advanced options helping you to know what content is indexed and how the extracted content is indexed. Throughout the book, you would go through complex problems with solutions along with varied approaches to tackle your business needs. By the end of this book, you will gain advanced proficiency to build out-of-box smart search solutions for your enterprise demands. What you will learn Design schema using schema API to access data in the database Advance querying and fine-tuning techniques for better performance Get to grips with indexing using Client API Set up a fault tolerant and highly available server with newer distributed capabilities, SolrCloud Explore Apache Tika to upload data with Solr Cell Understand different data operations that can be done while indexing Master advanced querying through Velocity Search UI, faceting and Query Re-ranking, pagination and spatial search Learn to use JavaScript, Python, SolrJ and Ruby for interacting with Solr Who this book is for The book would rightly appeal to developers, software engineers, data engineers and database architects who are building or seeking to build enterprise-wide effective search engines for business intelligence. Prior experience of Apache Solr or Java programming is must to take the best of this book.

This book describes the landscape of cloud computing from first principles, leading the reader step-by-step through the process of building and configuring a cloud environment. The book not only considers the technologies for designing and creating cloud computing platforms, but also the business models and frameworks in real-world implementation of cloud platforms. Emphasis is placed on “learning by doing,” and readers are encouraged to experiment with a range of different tools and approaches. Topics and features: includes review questions, hands-on exercises, study activities and discussion topics throughout the text; demonstrates the approaches used to build cloud computing

infrastructures; reviews the social, economic, and political aspects of the on-going growth in cloud computing use; discusses legal and security concerns in cloud computing; examines techniques for the appraisal of financial investment into cloud computing; identifies areas for further research within this rapidly-moving field.

How the Gifted Brain Learns

The Future of Email Archives

Principles and Practice

Managing and Processing Big Data in Cloud Computing

A Guide to Developing Next-Generation Enterprise Applications

Expert Hadoop 2 Administration

Data Lake for Enterprises

This volume includes 73 papers presented at ICTIS 2017: Second International Conference on Information and Communication Technology for Intelligent Systems. The conference was held on 25th and 26th March 2017, in Ahmedabad, India and organized jointly by the Associated Chambers of Commerce and Industry of India (ASSOCHAM) Gujarat Chapter, the G R Foundation, the Association of Computer Machinery, Ahmedabad Chapter and supported by the Computer Society of India Division IV – Communication and Division V – Education and Research. The papers featured mainly focus on information and communications technology (ICT) and its applications in intelligent computing, cloud storage, data mining and software analysis. The fundamentals of various data analytics and algorithms discussed are useful to researchers in the fiel d.

This book includes selected papers from the 2nd International Conference on Machine Learning and Information Processing (ICMLIP 2020), held at Vardhaman College of Engineering, Jawaharlal Nehru Technological University (JNTU), Hyderabad, India, from November 28 to 29, 2020. It presents the latest developments and technical solutions in the areas of advanced computing and data sciences, covering machine learning, artificial intelligence, human–computer interaction, IoT, deep learning, image processing and pattern recognition, and signal and speech processing.

Summary Tika in Action is a hands-on guide to content mining with Apache Tika. The book's many examples and case studies offer real-world experience from domains ranging from search engines to digital asset management and scientific data processing. About the Technology Tika is an Apache toolkit that has built into it everything you and your app need to know about file formats. Using Tika, your applications can discover and extract content from digital documents in almost any format, including exotic ones. About this Book Tika in Action is the ultimate guide to content mining using Apache Tika. You'll learn how to pull usable information from otherwise inaccessible sources,

including internet media and file archives. This example-rich book teaches you to build and extend applications based on real-world experience with search engines, digital asset management, and scientific data processing. In addition to architectural overviews, you'll find detailed chapters on features like metadata extraction, automatic language detection, and custom parser development. This book is written for developers who are new to both Scala and Lift and covers just enough Scala to get you started. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book. What's Inside Crack MS Word, PDF, HTML, and ZIP Integrate with search engines, CMS, and other data sources Learn through experimentation Many examples This book requires no previous knowledge of Tika or text mining techniques. It assumes a working knowledge of Java.

========== Table of Contents PART 1 GETTING STARTED The case for the digital Babel fish Getting started with Tika The information landscape PART 2 TIKA IN DETAIL Document type detection Content extraction Understanding metadata Language detection What's in a file? PART 3 INTEGRATION AND ADVANCED USE The big picture Tika and the Lucene search stack Extending Tika PART 4 CASE STUDIES Powering NASA science data systems Content management with Apache Jackrabbit Curating cancer research data with Tika The classic search engine example

Examine the techniques and Java tools supporting the growing field of data science About This Book Your entry ticket to the world of data science with the stability and power of Java Explore, analyse, and visualize your data effectively using easy-to-follow examples Make your Java applications more capable using machine learning Who This Book Is For This book is for Java developers who are comfortable developing applications in Java. Those who now want to enter the world of data science or wish to build intelligent applications will find this book ideal. Aspiring data scientists will also find this book very helpful. What You Will Learn Understand the nature and key concepts used in the field of data science Grasp how data is collected, cleaned, and processed Become comfortable with key data analysis techniques See specialized analysis techniques centered on machine learning Master the effective visualization of your data Work with the Java APIs and techniques used to perform data analysis In Detail Data science is concerned with extracting knowledge and insights from a wide variety of data sources to analyse patterns or predict future behaviour. It draws from a wide array of disciplines including statistics, computer science, mathematics, machine learning, and data mining. In this book, we cover the important data science concepts and how they are supported by Java, as

well as the often statistically challenging techniques, to provide you with an understanding of their purpose and application. The book starts with an introduction of data science, followed by the basic data science tasks of data collection, data cleaning, data analysis, and data visualization. This is followed by a discussion of statistical techniques and more advanced topics including machine learning, neural networks, and deep learning. The next section examines the major categories of data analysis including text, visual, and audio data, followed by a discussion of resources that support parallel implementation. The final chapter illustrates an in-depth data science problem and provides a comprehensive, Java-based solution. Due to the nature of the topic, simple examples of techniques are presented early followed by a more detailed treatment later in the book. This permits a more natural introduction to the techniques and concepts presented in the book. Style and approach This book follows a tutorial approach, providing examples of each of the major concepts covered. With a step-by-step instructional style, this book covers various facets of data science and will get you up and running quickly.

Information and Communication Technology for Intelligent Systems (ICTIS 2017) - Volume 2

Introducing Data Science

Logistics 4.0

Machine Learning with TensorFlow, Second Edition

Techniques for building machine learning and neural network models for NLP, 2nd Edition

Mahout in Action

A Practical Approach to Enterprise Search

Dig deep into the data with a hands-on guide to machine learning with updated examples and more! Machine Learning: Hands-On for Developers and Technical Professionals provides hands-on instruction and fully-coded working examples for the most common machine learning techniques used by developers and technical professionals. The book contains a breakdown of each ML variant, explaining how it works and how it is used within certain industries, allowing readers to incorporate the presented techniques into their own work as they follow along. A core tenant of machine learning is a strong focus on data preparation, and a full exploration of the various types of learning algorithms illustrates how the proper tools can help any developer extract information and insights from existing data. The book includes a full complement of Instructor's Materials to facilitate use in the classroom, making this resource useful for students and as a professional reference. At its core, machine learning is a mathematical, algorithm-based technology that forms the basis of historical data mining and modern big data science. Scientific analysis of big data requires a working knowledge of machine learning, which forms predictions based on known properties learned from training data. Machine Learning is an accessible, comprehensive guide for the non-mathematician, providing clear guidance that allows readers to: Learn the languages of machine learning including Hadoop, Mahout, and Weka Understand decision trees, Bayesian networks, and artificial neural networks Implement Association Rule, Real Time, and Batch learning Develop a strategic plan for safe, effective, and efficient machine learning

By learning to construct a system that can learn from data, readers can increase their utility across industries. Machine learning sits at the core of deep dive data analysis and visualization, which is increasingly in demand as companies discover the goldmine hiding in their existing data. For the tech professional involved in data science, Machine Learning: Hands-On for Developers and Technical Professionals provides the skills and techniques required to dig deeper.

R is the most powerful tool you can use for statistical analysis. This definitive guide smooths R's steep learning curve with practical solutions and real-world applications for commercial environments. In R in Action, Third Edition you will learn how to: Set up and install R and RStudio Clean, manage, and analyze data with R Use the ggplot2 package for graphs and visualizations Solve data management problems using R functions Fit and interpret regression models Test hypotheses and estimate confidence Simplify complex multivariate data with principal components and exploratory factor analysis Make predictions using time series forecasting Create dynamic reports and stunning visualizations Techniques for debugging programs and creating packages R in Action, Third Edition makes learning R quick and easy. That's why thousands of data scientists have chosen this guide to help them master the powerful language. Far from being a dry academic tome, every example you'll encounter in this book is relevant to scientific and business developers, and helps you solve common data challenges. R expert Rob Kabacoff takes you on a crash course in statistics, from dealing with messy and incomplete data to creating stunning visualizations. This revised and expanded third edition contains fresh coverage of the new tidyverse approach to data analysis and R's state-of-the-art graphing capabilities with the ggplot2 package. About the technology Used daily by data scientists, researchers, and quants of all types, R is the gold standard for statistical data analysis. This free and open source language includes packages for everything from advanced data visualization to deep learning. Instantly comfortable for mathematically minded users, R easily handles practical problems without forcing you to think like a software engineer. About the book R in Action, Third Edition teaches you how to do statistical analysis and data visualization using R and its popular tidyverse packages. In it, you'll investigate real-world data challenges, including forecasting, data mining, and dynamic report writing. This revised third edition adds new coverage for graphing with ggplot2, along with examples for machine learning topics like clustering, classification, and time series analysis. What's inside Clean, manage, and analyze data Use the ggplot2 package for graphs and visualizations Techniques for debugging programs and creating packages A complete learning resource for R and tidyverse About the reader Requires basic math and statistics. No prior experience with R needed. About the author Dr. Robert I Kabacoff is a professor of quantitative analytics at Wesleyan University and a seasoned data scientist with more than 20 years of experience. Table of Contents PART 1 GETTING STARTED 1 Introduction to R 2 Creating a dataset 3 Basic data management 4 Getting started with graphs 5 Advanced data management PART 2 BASIC METHODS 6 Basic graphs 7 Basic statistics PART 3 INTERMEDIATE METHODS 8 Regression 9 Analysis of variance 10 Power analysis 11 Intermediate graphs 12 Resampling statistics and bootstrapping PART 4 ADVANCED METHODS 13 Generalized linear models 14 Principal components and factor analysis 15 Time series 16 Cluster analysis 17 Classification 18 Advanced methods for missing data PART 5 EXPANDING YOUR SKILLS 19 Advanced graphs 20 Advanced programming 21 Creating dynamic reports 22 Creating a package

There's growing interest in learning how to analyze streaming data in large-scale systems such as web traffic, financial transactions, machine logs, industrial sensors, and many others. But analyzing data streams at scale has been difficult to do well—until now. This practical book delivers a deep introduction to Apache Flink, a highly innovative open source stream processor with a surprising range of capabilities. Authors Ellen Friedman and Kostas Tzoumas show technical and nontechnical readers alike how Flink is engineered to overcome significant tradeoffs that have limited the effectiveness of other approaches to stream processing. You'll also learn how Flink has the ability to handle both stream and batch

data processing with one technology. Learn the consequences of not doing streaming well—in retail and marketing, IoT, telecom, and banking and finance Explore how to design data architecture to gain the best advantage from stream processing Get an overview of Flink's capabilities and features, along with examples of how companies use Flink, including in production Take a technical dive into Flink, and learn how it handles time and stateful computation Examine how Flink processes both streaming (unbounded) and batch (bounded) data without sacrificing performance This is the eBook of the printed book and may not include any media, website access codes, or print supplements that may come packaged with the bound book. The Comprehensive, Up-to-Date Apache Hadoop Administration Handbook and Reference "Sam Alapati has worked with production Hadoop clusters for six years. His unique depth of experience has enabled him to write the go-to resource for all administrators looking to spec, size, expand, and secure production Hadoop clusters of any size." —Paul Dix, Series Editor In Expert Hadoop® Administration, leading Hadoop administrator Sam R. Alapati brings together authoritative knowledge for creating, configuring, securing, managing, and optimizing production Hadoop clusters in any environment. Drawing on his experience with large-scale Hadoop administration, Alapati integrates action-oriented advice with carefully researched explanations of both problems and solutions. He covers an unmatched range of topics and offers an unparalleled collection of realistic examples. Alapati demystifies complex Hadoop environments, helping you understand exactly what happens behind the scenes when you administer your cluster. You'll gain unprecedented insight as you walk through building clusters from scratch and configuring high availability, performance, security, encryption, and other key attributes. The high-value administration skills you learn here will be indispensable no matter what Hadoop distribution you use or what Hadoop applications you run. Understand Hadoop's architecture from an administrator's standpoint Create simple and fully distributed clusters Run MapReduce and Spark applications in a Hadoop cluster Manage and protect Hadoop data and high availability Work with HDFS commands, file permissions, and storage management Move data, and use YARN to allocate resources and schedule jobs Manage job workflows with Oozie and Hue Secure, monitor, log, and optimize Hadoop Benchmark and troubleshoot Hadoop

Concepts, Methodologies, Tools, and Applications

Dawn of the Code War

Building Digital Experience Platforms

XQuery Kick Start

Lucene in Action

Enhance Your Search with Faceted Navigation, Result Highlighting, Fuzzy Queries, Ranked Scoring, and More

Stream Processing for Real Time and Beyond

Tika in ActionSimon and Schuster

Summary Deep Learning for Search teaches you how to improve the effectiveness of your search by implementing neural network-based techniques. By the time you're finished with the book, you'll be ready to build amazing search engines that deliver the results your users need and that get better as time goes on! Foreword by Chris Mattmann. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Deep learning handles the toughest search challenges, including imprecise search terms, badly indexed data, and retrieving images with minimal metadata. And with modern tools like DL4J and TensorFlow, you can apply powerful DL techniques without a deep background in data science or natural language processing (NLP). This book will show you how. About the Book Deep Learning for Search teaches you to improve your search results

with neural networks. You'll review how DL relates to search basics like indexing and ranking. Then, you'll walk through in-depth examples to upgrade your search with DL techniques using Apache Lucene and Deeplearning4j. As the book progresses, you'll explore advanced topics like searching through images, translating user queries, and designing search engines that improve as they learn! What's inside Accurate and relevant rankings Searching across languages Content-based image search Search with recommendations About the Reader For developers comfortable with Java or a similar language and search basics. No experience with deep learning or NLP needed. About the Author Tommaso Teofili is a software engineer with a passion for open source and machine learning. As a member of the Apache Software Foundation, he contributes to a number of open source projects, ranging from topics like information retrieval (such as Lucene and Solr) to natural language processing and machine translation (including OpenNLP, Joshua, and UIMA). He currently works at Adobe, developing search and indexing infrastructure components, and researching the areas of natural language processing, information retrieval, and deep learning. He has presented search and machine learning talks at conferences including BerlinBuzzwords, International Conference on Computational Science, ApacheCon, EclipseCon, and others. You can find him on Twitter at @tteofili. Table of Contents PART 1 - SEARCH MEETS DEEP LEARNING Neural search Generating synonyms PART 2 - THROWING NEURAL NETS AT A SEARCH ENGINE From plain retrieval to text generation More-sensitive query suggestions Ranking search results with word embeddings Document embeddings for rankings and recommendations PART 3 - ONE STEP BEYOND Searching across languages Content-based image search A peek at performance

Summary Mahout in Action is a hands-on introduction to machine learning with Apache Mahout. Following real-world examples, the book presents practical use cases and then illustrates how Mahout can be applied to solve them. Includes a free audio- and video-enhanced ebook. About the Technology A computer system that learns and adapts as it collects data can be really powerful. Mahout, Apache's open source machine learning project, captures the core algorithms of recommendation systems, classification, and clustering in ready-to-use, scalable libraries. With Mahout, you can immediately apply to your own projects the machine learning techniques that drive Amazon, Netflix, and others. About this Book This book covers machine learning using Apache Mahout. Based on experience with real-world applications, it introduces practical use cases and illustrates how Mahout can be applied to solve them. It places particular focus on issues of scalability and how to apply these techniques against large data sets using the Apache Hadoop framework. This book is written for developers familiar with Java -- no prior experience with Mahout is assumed. Owners of a Manning pBook purchased anywhere in the world can download a free eBook from manning.com at any time. They can do so multiple times and in any or all formats available (PDF, ePub or Kindle). To do so, customers must register their printed copy on Manning's site by creating a user account and then following instructions printed on the pBook registration insert at the front of the book. What's Inside Use group data to make individual recommendations Find logical clusters within your data Filter and refine with on-the-fly classification Free audio and video extras Table of Contents Meet Apache Mahout PART 1 RECOMMENDATIONS Introducing recommenders Representing recommender data Making recommendations Taking recommenders to production Distributing recommendation computations PART 2 CLUSTERING Introduction to clustering Representing data Clustering algorithms in Mahout Evaluating and improving clustering quality Taking clustering to production Real-world applications of clustering PART 3 CLASSIFICATION Introduction to classification Training a classifier Evaluating and tuning a classifier Deploying a classifier Case study: Shop It To Me

The inside story of how America's enemies launched a cyber war against us-and how we've learned to fight back With each passing

year, the internet-linked attacks on America's interests have grown in both frequency and severity. Overmatched by our military, countries like North Korea, China, Iran, and Russia have found us vulnerable in cyberspace. The "Code War" is upon us. In this dramatic book, former Assistant Attorney General John P. Carlin takes readers to the front lines of a global but little-understood fight as the Justice Department and the FBI chases down hackers, online terrorist recruiters, and spies. Today, as our entire economy goes digital, from banking to manufacturing to transportation, the potential targets for our enemies multiply. This firsthand account is both a remarkable untold story and a warning of dangers yet to come.

Data analysis and graphics with R and Tidyverse

Java for Data Science

Machine Learning

Mastering Apache Solr 7.x

Natural Language Processing with Java

4th IFIP WG 5.5/SOCOLNET Doctoral Conference on Computing, Electrical and Industrial Systems, DoCEIS 2013, Costa de Caparica, Portugal, April 15-17, 2013, Proceedings

Apache Solr

Data in all domains is getting bigger. How can you work with it efficiently? Recently updated for Spark 1.3, this book introduces Apache Spark, the open source cluster computing system that makes data analytics fast to write and fast to run. With Spark, you can tackle big datasets quickly through simple APIs in Python, Java, and Scala. This edition includes new information on Spark SQL, Spark Streaming, setup, and Maven coordinates. Written by the developers of Spark, this book will have data scientists and engineers up and running in no time. You'll learn how to express parallel jobs with just a few lines of code, and cover applications from simple batch jobs to stream processing and machine learning. Quickly dive into Spark capabilities such as distributed datasets, in-memory caching, and the interactive shell Leverage Spark's powerful built-in libraries, including Spark SQL, Spark Streaming, and MLlib Use one programming paradigm instead of mixing and matching tools like Hive, Hadoop, Mahout, and Storm Learn how to deploy interactive, batch, and streaming applications Connect to data sources including HDFS, Hive, JSON, and S3 Master advanced topics like data partitioning and shared variables Updated with new code, new projects, and new chapters, Machine Learning with TensorFlow, Second Edition gives readers a solid foundation in machine-learning concepts and the TensorFlow library. Summary Updated with new code, new projects, and new chapters, Machine Learning with TensorFlow, Second Edition gives readers a solid foundation in machine-learning concepts and the TensorFlow library. Written by NASA JPL Deputy CTO and Principal Data Scientist Chris Mattmann, all examples are accompanied by downloadable Jupyter Notebooks for a hands-on experience coding TensorFlow with Python. New and revised content expands coverage of core machine learning algorithms, and advancements in neural networks such as VGG-Face facial identification classifiers and deep speech classifiers. Purchase of the print book

includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Supercharge your data analysis with machine learning! ML algorithms automatically improve as they process data, so results get better over time. You don't have to be a mathematician to use ML: Tools like Google's TensorFlow library help with complex calculations so you can focus on getting the answers you need. About the book Machine Learning with TensorFlow, Second Edition is a fully revised guide to building machine learning models using Python and TensorFlow. You'll apply core ML concepts to real-world challenges, such as sentiment analysis, text classification, and image recognition. Hands-on examples illustrate neural network techniques for deep speech processing, facial identification, and auto-encoding with CIFAR-10. What's inside Machine Learning with TensorFlow Choosing the best ML approaches Visualizing algorithms with TensorBoard Sharing results with collaborators Running models in Docker About the reader Requires intermediate Python skills and knowledge of general algebraic concepts like vectors and matrices. Examples use the super-stable 1.15.x branch of TensorFlow and TensorFlow 2.x. About the author Chris Mattmann is the Division Manager of the Artificial Intelligence, Analytics, and Innovation Organization at NASA Jet Propulsion Lab. The first edition of this book was written by Nishant Shukla with Kenneth Fricklas. Table of Contents PART 1 - YOUR MACHINE-LEARNING RIG 1 A machine-learning odyssey 2 TensorFlow essentials PART 2 - CORE LEARNING ALGORITHMS 3 Linear regression and beyond 4 Using regression for call-center volume prediction 5 A gentle introduction to classification 6 Sentiment classification: Large movie-review dataset 7 Automatically clustering data 8 Inferring user activity from Android accelerometer data 9 Hidden Markov models 10 Part-of-speech tagging and word-sense disambiguation PART 3 - THE NEURAL NETWORK PARADIGM 11 A peek into autoencoders 12 Applying autoencoders: The CIFAR-10 image dataset 13 Reinforcement learning 14 Convolutional neural networks 15 Building a real-world CNN: VGG-Face ad VGG-Face Lite 16 Recurrent neural networks 17 LSTMs and automatic speech recognition 18 Sequence-to-sequence models for chatbots 19 Utility landscape

Ready to unlock the power of your data? With this comprehensive guide, you'll learn how to build and maintain reliable, scalable, distributed systems with Apache Hadoop. This book is ideal for programmers looking to analyze datasets of any size, and for administrators who want to set up and run Hadoop clusters. You'll find illuminating case studies that demonstrate how Hadoop is used to solve specific problems. This third edition covers recent changes to Hadoop, including material on the new MapReduce API, as well as MapReduce 2 and its more flexible execution model (YARN). Store large datasets with the Hadoop Distributed File System (HDFS) Run distributed computations with MapReduce Use Hadoop's data and I/O building blocks for compression, data integrity, serialization (including Avro), and persistence Discover common pitfalls and advanced features for writing real-world MapReduce programs Design, build, and administer a dedicated Hadoop cluster—or run Hadoop in the cloud Load data from relational databases into HDFS, using Sqoop Perform large-scale data processing with the Pig query language Analyze datasets

with Hive, Hadoop’s data warehousing system Take advantage of HBase for structured and semi-structured data, and ZooKeeper for building distributed systems

It starts off by discussing Solr and helping you understand how it fits into your architecture_where all databases and document/web crawlers fall short, and Solr shines. The main part of the book is a thorough exploration of nearly every feature that Solr offers. To keep this interesting and realistic, we use a large open source set of metadata about artists, releases, and tracks courtesy of the MusicBrainz.org project. Using this data as a testing ground for Solr, you will learn how to import this data in various ways from CSV to XML to database access.

Apache Solr for Indexing Data

America's Battle Against Russia, China, and the Rising Global Cyber Threat

Practices, Techniques, and Applications

Social Big Data Analytics

Lightning-Fast Big Data Analysis

Big data, machine learning, and more, using Python tools

Hands-On for Developers and Technical Professionals

"XQuery Kick Start" delivers a concise introduction to the XQuery standard, and useful implementation advice for developers needing to put it into practice. The book starts by explaining the role of XQuery in the XML family of specifications, and its relationship with XPath. The authors then explain the specification in detail, describing the semantics and data model, before moving to examples using XQuery to manipulate XML databases and document storage systems. Later chapters discuss Java implementations of XQuery and development tools that facilitate the development of Web sites with XQuery. This book is up to date with the latest XQuery specifications, and includes coverage of new features for extending the XQuery language. This book covers three major parts of Big Data: concepts, theories and applications. Written by world-renowned leaders in Big Data, this book explores the problems, possible solutions and directions for Big Data in research and practice. It also focuses on high level concepts such as definitions of Big Data from different angles; surveys in research and applications; and existing tools, mechanisms, and systems in practice. Each chapter is independent from the other chapters, allowing users to read any chapter directly. After examining the practical side of Big Data, this book presents theoretical perspectives. The theoretical research ranges from Big Data representation, modeling and topology to distribution and dimension reducing. Chapters also investigate the many disciplines that involve Big Data, such as statistics, data mining,

machine learning, networking, algorithms, security and differential geometry. The last section of this book introduces Big Data applications from different communities, such as business, engineering and science. Big Data Concepts, Theories and Applications is designed as a reference for researchers and advanced level students in computer science, electrical engineering and mathematics. Practitioners who focus on information systems, big data, data mining, business analysis and other related fields will also find this material valuable.

Big data has presented a number of opportunities across industries. With these opportunities come a number of challenges associated with handling, analyzing, and storing large data sets. One solution to this challenge is cloud computing, which supports a massive storage and computation facility in order to accommodate big data processing. Managing and Processing Big Data in Cloud Computing explores the challenges of supporting big data processing and cloud-based platforms as a proposed solution. Emphasizing a number of crucial topics such as data analytics, wireless networks, mobile clouds, and machine learning, this publication meets the research needs of data analysts, IT professionals, researchers, graduate students, and educators in the areas of data science, computer programming, and IT development.

Identify, understand, and engage the full range of gifted learners with practical, brain-compatible classroom strategies! The updated edition of Sousa's bestseller translates the latest neuroscientific findings into practical strategies for engaging gifted and talented learners. Individual chapters are dedicated to talents in language, math, and the arts, and offer instructional applications for both elementary and secondary classrooms. This reader-friendly guide uncovers: How the brains of gifted students are different How to gauge if gifted students are being adequately challenged How to identify students who are both gifted and learning disabled How to better identify gifted minority students

Guide to Cloud Computing

A Report from the Task Force on Technical Approaches to Email Archives, July 2018

Real-time streaming applications with Apex

Solr in Action

Learning Apache Apex

Machine Learning and Information Processing

Solr 1.4 Enterprise Search Server

When Lucene first hit the scene five years ago, it was nothing short ofamazing. By using

this open-source, highly scalable, super-fast search engine,developers could integrate search into applications quickly and efficiently.A lot has changed since then-search has grown from a "nice-to-have" featureinto an indispensable part of most enterprise applications. Lucene now powerssearch in diverse companies including Akamai, Netflix, LinkedIn,Technorati, HotJobs, Epiphany, FedEx, Mayo Clinic, MIT, New ScientistMagazine, and many others. Some things remain the same, though. Lucene still delivers high-performancesearch features in a disarmingly easy-to-use API. Due to its vibrant and diverseopen-source community of developers and users, Lucene is relentlessly improving,with evolutions to APIs, significant new features such as payloads, and ahuge increase (as much as 8x) in indexing speed with Lucene 2.3. And with clear writing, reusable examples, and unmatched advice on bestpractices, Lucene in Action, Second Edition is still the definitive guide todeveloping with Lucene. Purchase of the print book comes with an offer of a free PDF, ePub, and Kindle eBook from Manning. Also available is all code from the book.

Industrial revolutions have impacted both, manufacturing and service. From the steam engine to digital automated production, the industrial revolutions have conduced significant changes in operations and supply chain management (SCM) processes. Swift changes in manufacturing and service systems have led to phenomenal improvements in productivity. The fast-paced environment brings new challenges and opportunities for the companies that are associated with the adaptation to the new concepts such as Internet of Things (IoT) and Cyber Physical Systems, artificial intelligence (AI), robotics, cyber security, data analytics, block chain and cloud technology. These emerging technologies facilitated and expedited the birth of Logistics 4.0. Industrial Revolution 4.0 initiatives in SCM has attracted stakeholders' attentions due to it is ability to empower using a set of technologies together that helps to execute more efficient production and distribution systems. This initiative has been called Logistics 4.0 of the fourth Industrial Revolution in SCM due to its high potential. Connecting entities, machines, physical items and enterprise resources to each other by using sensors, devices and the internet along the supply chains are the main attributes of Logistics 4.0. IoT enables

customers to make more suitable and valuable decisions due to the data-driven structure of the Industry 4.0 paradigm. Besides that, the system's ability of gathering and analyzing information about the environment at any given time and adapting itself to the rapid changes add significant value to the SCM processes. In this peer-reviewed book, experts from all over the world, in the field present a conceptual framework for Logistics 4.0 and provide examples for usage of Industry 4.0 tools in SCM. This book is a work that will be beneficial for both practitioners and students and academicians, as it covers the theoretical framework, on the one hand, and includes examples of practice and real world.

Summary Introducing Data Science teaches you how to accomplish the fundamental tasks that occupy data scientists. Using the Python language and common Python libraries, you'll experience firsthand the challenges of dealing with data at scale and gain a solid foundation in data science. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Many companies need developers with data science skills to work on projects ranging from social media marketing to machine learning. Discovering what you need to learn to begin a career as a data scientist can seem bewildering. This book is designed to help you get started. About the Book Introducing Data ScienceIntroducing Data Science explains vital data science concepts and teaches you how to accomplish the fundamental tasks that occupy data scientists. You'll explore data visualization, graph databases, the use of NoSQL, and the data science process. You'll use the Python language and common Python libraries as you experience firsthand the challenges of dealing with data at scale. Discover how Python allows you to gain insights from data sets so big that they need to be stored on multiple machines, or from data moving so quickly that no single machine can handle it. This book gives you hands-on experience with the most popular Python data science libraries, Scikit-learn and StatsModels. After reading this book, you'll have the solid foundation you need to start a career in data science. What's Inside Handling large data Introduction to machine learning Using Python to work with data Writing data science algorithms About the Reader This book assumes you're comfortable reading code in Python or a similar language,

such as C, Ruby, or JavaScript. No prior experience with data science is required. About the Authors Davy Cielen, Arno D. B. Meysman, and Mohamed Ali are the founders and managing partners of Optimately and Maiton, where they focus on developing data science projects and solutions in various sectors. Table of Contents Data science in a big data world The data science process Machine learning Handling large data on a single computer First steps in big data Join the NoSQL movement The rise of graph databases Text mining and text analytics Data visualization to the end user

Designing and writing a real-time streaming publication with Apache Apex About This Book Get a clear, practical approach to real-time data processing Program Apache Apex streaming applications This book shows you Apex integration with the open source Big Data ecosystem Who This Book Is For This book assumes knowledge of application development with Java and familiarity with distributed systems. Familiarity with other real-time streaming frameworks is not required, but some practical experience with other big data processing utilities might be helpful. What You Will Learn Put together a functioning Apex application from scratch Scale an Apex application and configure it for optimal performance Understand how to deal with failures via the fault tolerance features of the platform Use Apex via other frameworks such as Beam Understand the DevOps implications of deploying Apex In Detail Apache Apex is a next-generation stream processing framework designed to operate on data at large scale, with minimum latency, maximum reliability, and strict correctness guarantees. Half of the book consists of Apex applications, showing you key aspects of data processing pipelines such as connectors for sources and sinks, and common data transformations. The other half of the book is evenly split into explaining the Apex framework, and tuning, testing, and scaling Apex applications. Much of our economic world depends on growing streams of data, such as social media feeds, financial records, data from mobile devices, sensors and machines (the Internet of Things - IoT). The projects in the book show how to process such streams to gain valuable, timely, and actionable insights. Traditional use cases, such as ETL, that currently consume a significant chunk of data engineering resources are also covered. The final chapter shows you future possibilities emerging in the streaming space, and how Apache

Apex can contribute to it. Style and approach This book is divided into two major parts: first it explains what Apex is, what its relevant parts are, and how to write well-built Apex applications. The second part is entirely application-driven, walking you through Apex applications of increasing complexity.

Big Data Concepts, Theories, and Applications

Managing Spark, YARN, and MapReduce

Taming Text

CMIS and Apache Chemistry in Action

Tika in Action

An expert guide to advancing, optimizing, and scaling your enterprise search

Use digital experience platforms (DXP) to improve your development productivity and release timelines. Leverage the pre-integrated feature sets of DXPs in your organization's digital transformation journey to quickly develop a personalized, secure, and robust enterprise platform. In this book the authors examine various features of DXPs and provide rich insights into building each layer in a digital platform. Proven best practices are presented with examples for designing and building layers. A special focus is provided on security and quality attributes needed for business-critical enterprise applications. The authors cover modern and emerging digital trends such as Blockchain, IoT, containers, chatbots, artificial intelligence, and more. The book is divided into five parts related to requirements/design, development, security, infrastructure, and case study. The authors employ proven real-world methods, best practices, and security and integration techniques derived from their rich experience. An elaborate digital transformation case study for a banking application is included. What You'll Learn Develop a digital experience platform from end to endUnderstand best practices and proven methods for designing overall architecture, user interface and integration components, security, and infrastructureStudy real-world cases, including an elaborate digital transformation building an enterprise platform for a banking applicationKnow the open source tools and technology frameworks that can be used to build DXPs Who This Book Is For Web developers, full stack developers, digital enthusiasts, digital project managers, and architects Build an enterprise search engine using Apache Solr: index and search documents; ingest data from varied sources; apply various text processing techniques; utilize different search capabilities; and customize Solr to retrieve the desired results. Apache Solr: A Practical Approach to Enterprise Search explains each essential concept-backed by practical and industry examples--to help you attain expert-level knowledge. The book, which assumes a basic knowledge of Java, starts with an introduction to Solr, followed by steps to setting it up, indexing your first set of documents, and searching them. It then introduces you to information retrieval and its implementation in Apache Solr; this will help you understand your search problem, decide the approach to build

an effective solution, and use various metrics to evaluate the results. The book next covers the schema design and techniques to build a text analysis chain for cleansing, normalizing and enriching your documents and addressing different types of search queries. It describes various popular matching techniques which are generally applied to improve the precision and recall of searches. You will learn the end-to-end process of data ingestion from varied sources, metadata extraction, pre-processing and transformation of content, various search components, query parsers and other advanced search capabilities. After covering out-of-the-box features, Solr expert Dikshant Shahi dives into ways you can customize Solr for your business and its specific requirements, along with ways to plug in your own components. Most important, you will learn about implementations for Solr scoring, factors affecting the document score, and tuning the score for the application at hand. The book explains why textual scoring is not sufficient for practical ranking of documents and ways to integrate real-world factors for contributing to the document ranking. You'll see how to influence user experience by providing suggestions and recommendations. You'll also see integration of Solr with important related technologies such as OpenNLP and Tika. Additionally, you will learn about scaling Solr using SolrCloud. This book concludes with coverage of semantic search capabilities, which is crucial for taking the search experience to the next level. By the end of Apache Solr, you will be proficient in designing and developing your search engine. 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 t 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 a 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 int 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 e 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.