

Apache Solr A Practical Approach To Enterprise Search

Thank you very much for reading **Apache Solr A Practical Approach To Enterprise Search** . Maybe you have knowledge that, people have look numerous times for their favorite readings like this Apache Solr A Practical Approach To Enterprise Search , but end up in malicious downloads.

Rather than reading a good book with a cup of tea in the afternoon, instead they are facing with some harmful bugs inside their desktop computer.

Apache Solr A Practical Approach To Enterprise Search is available in our book collection an online access to it is set as public so you can download it instantly.

Our books collection spans in multiple countries, allowing you to get the most less latency time to download any of our books like this one.

Merely said, the Apache Solr A Practical Approach To Enterprise Search is universally compatible with any devices to read

Scaling Apache Solr - Hrishikesh Vijay Karambelkar 2014-07-25

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.

[Apache Kafka 1.0 Cookbook](#) - Raúl Estrada 2017-12-21

Simplify real-time data processing by leveraging the power of Apache Kafka 1.0 Key Features Use Kafka 1.0 features such as Confluent platforms and Kafka streams to build efficient streaming data applications to handle and process your data Integrate Kafka with other Big Data tools such as Apache Hadoop, Apache Spark, and more Hands-on recipes to help you design, operate, maintain, and secure your Apache Kafka cluster with ease Book Description Apache Kafka provides a unified, high-throughput, low-latency platform to handle real-time data feeds. This book will show you how to use Kafka efficiently, and contains practical solutions

to the common problems that developers and administrators usually face while working with it. This practical guide contains easy-to-follow recipes to help you set up, configure, and use Apache Kafka in the best possible manner. You will use Apache Kafka Consumers and Producers to build effective real-time streaming applications. The book covers the recently released Kafka version 1.0, the Confluent Platform and Kafka Streams. The programming aspect covered in the book will teach you how to perform important tasks such as message validation, enrichment and composition. Recipes focusing on optimizing the performance of your Kafka cluster, and integrate Kafka with a variety of third-party tools such as Apache Hadoop, Apache Spark, and Elasticsearch will help ease your day to day collaboration with Kafka greatly. Finally, we cover tasks related to monitoring and securing your Apache Kafka cluster using tools such as Ganglia and Graphite. If you're looking to become the go-to person in your organization when it comes to working with Apache Kafka, this book is the only resource you need to have. What you will learn -Install and configure Apache Kafka 1.0 to get optimal performance -

Create and configure Kafka Producers and Consumers -Operate your Kafka clusters efficiently by implementing the mirroring technique -Work with the new Confluent platform and Kafka streams, and achieve high availability with Kafka -Monitor Kafka using tools such as Graphite and Ganglia - Integrate Kafka with third-party tools such as Elasticsearch, Logstash, Apache Hadoop, Apache Spark, and more Who this book is for This book is for developers and Kafka administrators who are looking for quick, practical solutions to problems encountered while operating, managing or monitoring Apache Kafka. If you are a developer, some knowledge of Scala or Java will help, while for administrators, some working knowledge of Kafka will be useful.

Open Source Software in Life Science Research - Lee Harland 2012-10-31

The free/open source approach has grown from a minor activity to become a significant producer of robust, task-orientated software for a wide variety of situations and applications. To life science informatics groups, these systems present an appealing proposition - high quality software at a very attractive price. Open source software in life science research considers how industry and applied research groups have embraced these resources, discussing practical implementations that address real-world business problems. The book is divided into four parts. Part one looks at laboratory data management and chemical informatics, covering software such as Bioclipse, OpenTox, ImageJ and KNIME. In part two, the focus turns to genomics and bioinformatics tools, with chapters examining GenomicsTools and EBI Atlas software, as well as the practicalities of setting up an 'omics' platform and managing large volumes of data. Chapters in part three examine information and knowledge management, covering a range of topics including software for web-based collaboration, open source search and visualisation technologies for scientific business applications, and specific software such as DesignTracker and Utopia

Documents. Part four looks at semantic technologies such as Semantic MediaWiki, TripleMap and Chem2Bio2RDF, before part five examines clinical analytics, and validation and regulatory compliance of free/open source software. Finally, the book concludes by looking at future perspectives and the economics and free/open source software in industry. Discusses a broad range of applications from a variety of sectors Provides a unique perspective on work normally performed behind closed doors Highlights the criteria used to compare and assess different approaches to solving problems

Learning Apache Mahout Classification

- Ashish Gupta 2015-02-26

If you are a data scientist who has some experience with the Hadoop ecosystem and machine learning methods and want to try out classification on large datasets using Mahout, this book is ideal for you.

Knowledge of Java is essential.

Elasticsearch: The Definitive Guide -

Clinton Gormley 2015-01-23

Whether you need full-text search or real-time analytics of structured data—or both—the Elasticsearch distributed search engine is an ideal way to put your data to work. This practical guide not only shows you how to search, analyze, and explore data with Elasticsearch, but also helps you deal with the complexities of human language, geolocation, and relationships. If you're a newcomer to both search and distributed systems, you'll quickly learn how to integrate Elasticsearch into your application. More experienced users will pick up lots of advanced techniques. Throughout the book, you'll follow a problem-based approach to learn why, when, and how to use Elasticsearch features. Understand how Elasticsearch interprets data in your documents Index and query your data to take advantage of search concepts such as relevance and word proximity Handle human language through the effective use of analyzers and queries Summarize and group data to show overall trends, with aggregations and analytics Use geo-points and geo-

shapes—Elasticsearch’s approaches to geolocation Model your data to take advantage of Elasticsearch’s horizontal scalability Learn how to configure and monitor your cluster in production

Cassandra: The Definitive Guide - Jeff Carpenter 2016-06-29

Imagine what you could do if scalability wasn't a problem. With this hands-on guide, you'll learn how the Cassandra database management system handles hundreds of terabytes of data while remaining highly available across multiple data centers. This expanded second edition—updated for Cassandra 3.0—provides the technical details and practical examples you need to put this database to work in a production environment. Authors Jeff Carpenter and Eben Hewitt demonstrate the advantages of Cassandra’s non-relational design, with special attention to data modeling. If you’re a developer, DBA, or application architect looking to solve a database scaling issue or future-proof your application, this guide helps you harness Cassandra’s speed and flexibility. Understand Cassandra’s distributed and decentralized structure Use the Cassandra Query Language (CQL) and cqlsh—the CQL shell Create a working data model and compare it with an equivalent relational model Develop sample applications using client drivers for languages including Java, Python, and Node.js Explore cluster topology and learn how nodes exchange data Maintain a high level of performance in your cluster Deploy Cassandra on site, in the Cloud, or with Docker Integrate Cassandra with Spark, Hadoop, Elasticsearch, Solr, and Lucene

Practical Natural Language Processing - Sowmya Vajjala 2020-06-17

Many books and courses tackle natural language processing (NLP) problems with toy use cases and well-defined datasets. But if you want to build, iterate, and scale NLP systems in a business setting and tailor them for particular industry verticals, this is your guide. Software engineers and data scientists will learn how to navigate the maze of options available at each step of the journey. Through the course of the

book, authors Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, and Harshit Surana will guide you through the process of building real-world NLP solutions

embedded in larger product setups. You'll learn how to adapt your solutions for different industry verticals such as healthcare, social media, and retail. With this book, you'll: Understand the wide spectrum of problem statements, tasks, and solution approaches within NLP Implement and evaluate different NLP applications using machine learning and deep learning methods Fine-tune your NLP solution based on your business problem and industry vertical Evaluate various algorithms and approaches for NLP product tasks, datasets, and stages Produce software solutions following best practices around release, deployment, and DevOps for NLP systems Understand best practices, opportunities, and the roadmap for NLP from a business and product leader’s perspective

Next-Generation Big Data - Butch Quinto 2018-06-12

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

examples, and practical advice Integrate HBase, Solr, Oracle, SQL Server, MySQL, Flume, Kafka, HDFS, and Amazon S3 with Apache Kudu, Impala, and Spark Use StreamSets, Talend, Pentaho, and CDAP for real-time and batch data ingestion and processing Utilize Trifacta, Alteryx, and Datameer for data wrangling and interactive data processing Turbocharge Spark with Alluxio, a distributed in-memory storage platform Deploy big data in the cloud using Cloudera Director Perform real-time data visualization and time series analysis using Zoomdata, Apache Kudu, Impala, and Spark Understand enterprise big data topics such as big data governance, metadata management, data lineage, impact analysis, and policy enforcement, and how to use Cloudera Navigator to perform common data governance tasks Implement big data use cases such as big data warehousing, data warehouse optimization, Internet of Things, real-time data ingestion and analytics, complex event processing, and scalable predictive modeling Study real-world big data case studies from innovative companies, including Navistar, Cerner, British Telecom, Shopzilla, Thomson Reuters, and Mastercard Who This Book Is For BI and big data warehouse professionals interested in gaining practical and real-world insight into next-generation big data processing and analytics using Apache Kudu, Impala, and Spark; and those who want to learn more about other advanced enterprise topics

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

A practical guide to implementing your enterprise data lake using Lambda Architecture as the base About This Book Build a full-fledged data lake for your organization with popular big data technologies using the Lambda architecture as the base Delve into the big data technologies required to meet modern day business strategies A highly practical guide to implementing enterprise data lakes with lots of examples and real-world use-cases Who This Book Is For Java developers and

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

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

8th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2014)

- Julio Saez-Rodriguez 2014-05-21
Biological and biomedical research are increasingly driven by experimental techniques that challenge our ability to analyse, process and extract meaningful knowledge from the underlying data. The impressive capabilities of next generation sequencing technologies, together with novel and ever evolving distinct types of omics data technologies, have put an increasingly complex set of challenges for the growing fields of Bioinformatics and Computational Biology. The analysis of the datasets produced and their integration call for new algorithms and approaches from fields such as Databases, Statistics, Data Mining, Machine Learning, Optimization, Computer Science and Artificial Intelligence. Clearly, Biology is more and more a science of information requiring tools from the computational sciences. In the last few years, we have seen the surge of a new generation of interdisciplinary scientists that have a strong background in the biological and computational sciences. In this context, the interaction of researchers from different scientific fields is, more than ever, of foremost importance boosting the research efforts in the field and contributing to the education of a new generation of Bioinformatics scientists. PACBB'14 contributes to this effort

promoting this fruitful interaction. PACBB'14 technical program included 34 papers spanning many different sub-fields in Bioinformatics and Computational Biology. Therefore, the conference promotes the interaction of scientists from diverse research groups and with a distinct background such as computer scientists, mathematicians or biologists.

Practical Digital Preservation - Adrian Brown 2013-05-23

A practical guide to the development and operation of digital preservation services for organizations of any size Practical Digital Preservation offers a comprehensive overview of best practice and is aimed at the non-specialist, assuming only a basic understanding of IT. The book provides guidance as to how to implement strategies with minimal time and resources. Digital preservation has become a critical issue for institutions of all sizes but until recently has mostly been the preserve of national archives and libraries with the resources, time and specialist knowledge available to experiment. As the discipline matures and practical tools and information are increasingly available the barriers to entry are falling for smaller organizations which can realistically start to take active steps towards a preservation strategy. However, the sheer volume of technical information now available on the subject is becoming a significant obstacle and a straightforward guide is required to offer clear and practical solutions. Each chapter in Practical Digital Preservation covers the essential building blocks of digital preservation strategy and implementation, leading the reader through the process. International case studies from organizations such as the Wellcome Library, Central Connecticut State University Library in the USA and Gloucestershire Archives in the UK illustrate how real organizations have approached the challenges of digital preservation. Key topics include: • Making the case for digital preservation • Understanding your requirements • Models for implementing a digital preservation

service • Selecting and acquiring digital objects • Accessioning and ingesting digital objects • Describing digital objects • Preserving digital objects • Providing access to users • Future trends.

Readership: Anyone involved in digital preservation and those wanting to get a better understanding of the process, students studying library and information science (LIS), archives and records management courses and academics getting to grips with practical issues.

Apache Solr PHP Integration - Jayant Kumar 2013-11-25

This book is full of step-by-step example-oriented tutorials which will show readers how to integrate Solr in PHP applications using the available libraries, and boost the inherent search facilities that Solr offers. If you are a developer who knows PHP and is interested in integrating search into your applications, this is the book for you. No advanced knowledge of Solr is required. Very basic knowledge of system commands and the command-line interface on both Linux and Windows is required. You should also be familiar with the concept of Web servers.

Transforming the IT Services Lifecycle with AI Technologies - Kristof Kloeckner 2018-09-20

As more and more industries are experiencing digital disruption, using information technology to enable a competitive advantage becomes a critical success factor for all enterprises. This book covers the authors' insights on how AI technologies can fundamentally reshape the IT services delivery lifecycle to deliver better business outcomes through a data-driven and knowledge-based approach. Three main challenges and the technologies to address them are discussed in detail:

- Gaining actionable insight from operational data for service management automation and improved human decision making
- Capturing and enhancing expert knowledge throughout the lifecycle from solution design to ongoing service improvement
- Enabling self-service for service requests and problem resolution, through intuitive

natural language interfaces The authors are top researchers and practitioners with deep experience in the fields of artificial intelligence and IT service management and are discussing both practical advice for IT teams and advanced research results.

The topics appeal to CIOs and CTOs as well as researchers who want to understand the state of the art of applying artificial intelligence to a very complex problem space. Although the book is concise, it comprehensively discuss topics like gaining insight from operational data for automatic problem diagnosis and resolution as well as continuous service optimization, AI for solution design and conversational self-service systems.

Apache Solr Enterprise Search Server - Third Edition - David Smiley 2015-05-26

This book is for developers who want to learn how to get the most out of Solr in their applications, whether you are new to the field, have used Solr but don't know everything, or simply want a good reference. It would be helpful to have some familiarity with basic programming concepts, but no prior experience is required.

Mastering Apache Solr - Mr. Mathieu Nayrolles 2014-05-15

Topic: In the open source, full-text search community, a leader emerges - Apache Solr. Apache Solr enables you to index and access documents orders of magnitude faster than classical databases and thereby provides a first-class search experience to your end users. Brief Description:

Mastering Apache Solr is a practical, hands-on guide containing crisp, relevant, systematically arranged, and progressive chapters. These chapters contain a wealth of information presented in a direct and easy-to-understand manner. This book covers key technical concepts, highlighting Solr's supremacy over classical databases in full-text search, which will help you accelerate your progress in the Solr world. Detailed Description: Mastering Apache Solr starts with an introduction to Apache Solr, its underlying technologies, the main differences between the classical database

engines, and gradually moves to more advance topics like boosting performance. In this book, we will look under the hood of a large number of topics and discuss answers to pertinent questions like why denormalize data, how to import classical databases' data inside Apache Solr, how to serve Solr through five different web servers, how to optimize them to serve Solr even faster. An important and major topic covered in this book is Solr's querying mechanism, which will prove to be a strongly in our journey through this book. We then look at boosting performance and deploying Solr using several servlet servers. Finally, we cover how to communicate with Solr using different programming languages, before deploying it in a cloud-based environment. Who this book is for: Mastering Apache Solr has been written for developers, programmers, and data specialists who want to take a leap towards the future of full-text storage and search and offer a world-class experience to their users. The reader is expected to have a working knowledge of traditional databases, Linux-based operating systems, and XML configuration files. Style and Approach: Mastering Apache Solr is written lucidly and has a dynamically simple approach. From the first page to the last, the book remains practical and focuses on the most important topics used in the world of Apache Solr without neglecting important theoretical fundamentals that help you build a strong foundation. Conclusion: Mastering Apache Solr will empower you to provide a world-class search experience to your end users through the discovery of the powerful mechanisms presented in this book.

[Scaling Big Data with Hadoop and Solr - Second Edition](#) - Hrishikesh Vijay Karambelkar 2015-04-27

This book is aimed at developers, designers, and architects who would like to build big data enterprise search solutions for their customers or organizations. No prior knowledge of Apache Hadoop and Apache Solr/Lucene technologies is required.

Solr in Action - Trey Grainger 2014-04-05

Summary Solr in Action is a comprehensive guide to implementing scalable search using Apache Solr. This clearly written book walks you through well-documented examples ranging from basic keyword searching to scaling a system for billions of documents and queries. It will give you a deep understanding of how to implement core Solr capabilities. About the Book Whether you're handling big (or small) data, managing documents, or building a website, it is important to be able to quickly search through your content and discover meaning in it. Apache Solr is your tool: a ready-to-deploy, Lucene-based, open source, full-text search engine. Solr can scale across many servers to enable real-time queries and data analytics across billions of documents. Solr in Action teaches you to implement scalable search using Apache Solr. This easy-to-read guide balances conceptual discussions with practical examples to show you how to implement all of Solr's core capabilities. You'll master topics like text analysis, faceted search, hit highlighting, result grouping, query suggestions, multilingual search, advanced geospatial and data operations, and relevancy tuning. This book assumes basic knowledge of Java and standard database technology. No prior knowledge of Solr or Lucene is required. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. What's Inside How to scale Solr for big data Rich real-world examples Solr as a NoSQL data store Advanced multilingual, data, and relevancy tricks Coverage of versions through Solr 4.7 About the Authors Trey Grainger is a director of engineering at CareerBuilder. Timothy Potter is a senior member of the engineering team at LucidWorks. The authors work on the scalability and reliability of Solr, as well as on recommendation engine and big data analytics technologies. Table of Contents PART 1 MEET SOLR Introduction to Solr Getting to know Solr Key Solr concepts Configuring Solr Indexing Text analysis PART 2 CORE SOLR CAPABILITIES

Performing queries and handling results
Faceted search Hit highlighting Query
suggestions Result grouping/field collapsing
Taking Solr to production PART 3 TAKING
SOLR TO THE NEXT LEVEL SolrCloud
Multilingual search Complex query
operations Mastering relevancy

Taming Text - Grant Ingersoll 2012-12-20
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 drowning in it.
Fortunately, there are innovative tools and
techniques for managing unstructured
information that can throw the smart
developer a much-needed lifeline. You'll
find them in this book. Taming Text is a
practical, example-driven guide to working
with text 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 the foundations upon which they are
built. Written in a clear and concise style,
this book avoids jargon, explaining the
subject in terms you can understand
without a background in statistics or natural
language processing. Examples are in 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 Lucene committer, 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 very complex
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

Apache Pulsar in Action - David
Kjerrumgaard 2021-12-28
Deliver lightning fast and reliable
messaging for your distributed applications
with the flexible and resilient Apache Pulsar
platform. In Apache Pulsar in Action you
will learn how to: Publish from Apache
Pulsar into third-party data repositories and
platforms Design and develop Apache
Pulsar functions Perform interactive SQL
queries against data stored in Apache
Pulsar Apache Pulsar in Action is a
comprehensive and practical guide to
building high-traffic applications with
Pulsar. You'll learn to use this mature and
battle-tested platform to deliver extreme
levels of speed and durability to your
messaging. Apache Pulsar committer David
Kjerrumgaard teaches you to apply Pulsar's
seamless scalability through hands-on case
studies, including IOT analytics applications
and a microservices app based on Pulsar
functions. Purchase of the print book
includes a free eBook in PDF, Kindle, and
ePub formats from Manning Publications.
About the technology Reliable server-to-
server messaging is the heart of a
distributed application. Apache Pulsar is a
flexible real-time messaging platform built
to run on Kubernetes and deliver the

scalability and resilience required for cloud-based systems. Pulsar supports both streaming and message queuing, and unlike other solutions, it can communicate over multiple protocols including MQTT, AMQP, and Kafka's binary protocol. About the book Apache Pulsar in Action teaches you to build scalable streaming messaging systems using Pulsar. You'll start with a rapid introduction to enterprise messaging and discover the unique benefits of Pulsar. Following crystal-clear explanations and engaging examples, you'll use the Pulsar Functions framework to develop a microservices-based application. Real-world case studies illustrate how to implement the most important messaging design patterns. What's inside Publish from Pulsar into third-party data repositories and platforms Design and develop Apache Pulsar functions Create an event-driven food delivery application About the reader Written for experienced Java developers. No prior knowledge of Pulsar required. About the author David Kjerrumgaard is a committer on the Apache Pulsar project. He currently serves as a Developer Advocate for StreamNative, where he develops Pulsar best practices and solutions. Table of Contents PART 1 GETTING STARTED WITH APACHE PULSAR 1 Introduction to Apache Pulsar 2 Pulsar concepts and architecture 3 Interacting with Pulsar PART 2 APACHE PULSAR DEVELOPMENT ESSENTIALS 4 Pulsar functions 5 Pulsar IO connectors 6 Pulsar security 7 Schema registry PART 3 HANDS-ON APPLICATION DEVELOPMENT WITH APACHE PULSAR 8 Pulsar Functions patterns 9 Resiliency patterns 10 Data access 11 Machine learning in Pulsar 12 Edge analytics

Introduction to Information Retrieval -

Christopher D. Manning 2008-07-07

Class-tested and coherent, this textbook teaches classical and web information retrieval, including web search and the related areas of text classification and text clustering from basic concepts. It gives an up-to-date treatment of all aspects of the design and implementation of systems for gathering, indexing, and searching

documents; methods for evaluating systems; and an introduction to the use of machine learning methods on text collections. All the important ideas are explained using examples and figures, making it perfect for introductory courses in information retrieval for advanced undergraduates and graduate students in computer science. Based on feedback from extensive classroom experience, the book has been carefully structured in order to make teaching more natural and effective. Slides and additional exercises (with solutions for lecturers) are also available through the book's supporting website to help course instructors prepare their lectures.

Developing RESTful Web Services with

Jersey 2.0 - Sunil Gulabani 2013-11

The book will follow a standard tutorial approach and will teach readers how to use the Jersey API for creating RESTful web services. This book is intended for Java EE developers who are building applications on the REST architecture. This is a quick, hands-on guide for learning JAX-RS 2.0. Developers should have some knowledge about RESTful web services but it's not essential to know JAX-RS 1.0.

Mastering Apache Solr 7.x - Sandeep Nair 2018-02-22

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.

Practical Cassandra - Russell Bradberry 2013

"Eric and Russell were early adopters of Cassandra at SimpleReach. In Practical

Cassandra, you benefit from their experience in the trenches administering Cassandra, developing against it, and building one of the first CQL drivers. If you are deploying Cassandra soon, or you inherited a Cassandra cluster to tend, spend some time with the deployment, performance tuning, and maintenance chapters... If you are new to Cassandra, I highly recommend the chapters on data modeling and CQL." –From the Foreword by Jonathon Ellis, Apache Cassandra Chair Build and Deploy Massively Scalable, Super-fast Data Management Applications with Apache Cassandra Practical Cassandra is the first hands-on developer's guide to building Cassandra systems and applications that deliver breakthrough speed, scalability, reliability, and performance. Fully up to date, it reflects the latest versions of Cassandra—including Cassandra Query Language (CQL), which dramatically lowers the learning curve for Cassandra developers. Pioneering Cassandra developers and Datastax MVPs Russell Bradberry and Eric Lubow walk you through every step of building a real production application that can store enormous amounts of structured, semi-structured, and unstructured data. Drawing on their exceptional expertise, Bradberry and Lubow share practical insights into issues ranging from querying to deployment, management, maintenance, monitoring, and troubleshooting. The authors cover key issues, from architecture to migration, and guide you through crucial decisions about configuration and data modeling. They provide tested sample code, detailed explanations of how Cassandra works "under the covers," and new case studies from three cutting-edge users: Ooyala, Hailo, and eBay. Coverage includes Understanding Cassandra's approach, architecture, key concepts, and primary use cases- and why it's so blazingly fast Getting Cassandra up and running on single nodes and large clusters Applying the new design patterns, philosophies, and features that make Cassandra such a powerful data store Leveraging CQL to simplify your transition

from SQL-based RDBMSes Deploying and provisioning through the cloud or on bare-metal hardware Choosing the right configuration options for each type of workload Tweaking Cassandra to get maximum performance from your hardware, OS, and JVM Mastering Cassandra's essential tools for maintenance and monitoring Efficiently solving the most common problems with Cassandra deployment, operation, and application development

[Apache Solr Search Patterns](#) - Jayant Kumar 2015-04-24

This book is for developers who already know how to use Solr and are looking at procuring advanced strategies for improving their search using Solr. This book is also for people who work with analytics to generate graphs and reports using Solr. Moreover, if you are a search architect who is looking forward to scale your search using Solr, this is a must have book for you. It would be helpful if you are familiar with the Java programming language.

[Microservices and Containers](#) - Parminder Singh Kocher 2018-03-16

Transition to Microservices and DevOps to Transform Your Software Development Effectiveness Thanks to the tech sector's latest game-changing innovations—the Internet of Things (IoT), software-enabled networking, and software as a service (SaaS), to name a few—there is now a seemingly insatiable demand for platforms and architectures that can improve the process of application development and deployment. In *Microservices and Containers*, longtime systems architect and engineering team leader Parminder Kocher analyzes two of the hottest new technology trends: microservices and containers. Together, as Kocher demonstrates, microservices and Docker containers can bring unprecedented agility and scalability to application development and deployment, especially in large, complex projects where speed is crucial but small errors can be disastrous. Learn how to leverage microservices and Docker to drive

modular architectural design, on-demand scalability, application performance and reliability, time-to-market, code reuse, and exponential improvements in DevOps effectiveness. Kocher offers detailed guidance and a complete roadmap for transitioning from monolithic architectures, as well as an in-depth case study that walks the reader through the migration of an enterprise-class SOA system. Understand how microservices enable you to organize applications into standalone components that are easier to manage, update, and scale Decide whether microservices and containers are worth your investment, and manage the organizational learning curve associated with them Apply best practices for interprocess communication among microservices Migrate monolithic systems in an orderly fashion Understand Docker containers, installation, and interfaces Network, orchestrate, and manage Docker containers effectively Use Docker to maximize scalability in microservices-based applications Apply your learning with an in-depth, hands-on case study Whether you are a software architect/developer or systems professional looking to move on from older approaches or a manager trying to maximize the business value of these technologies, *Microservices and Containers* will be an invaluable addition to your library. Register your product at informit.com/register for convenient access to downloads, updates, and/or corrections as they become available.

Advanced Elasticsearch 7.0 - Wai Tak Wong 2019-08-23

Master the intricacies of Elasticsearch 7.0 and use it to create flexible and scalable search solutions Key Features Master the latest distributed search and analytics capabilities of Elasticsearch 7.0 Perform searching, indexing, and aggregation of your data at scale Discover tips and techniques for speeding up your search query performance Book Description Building enterprise-grade distributed applications and executing systematic search operations call for a strong understanding of Elasticsearch and

expertise in using its core APIs and latest features. This book will help you master the advanced functionalities of Elasticsearch and understand how you can develop a sophisticated, real-time search engine confidently. In addition to this, you'll also learn to run machine learning jobs in Elasticsearch to speed up routine tasks. You'll get started by learning to use Elasticsearch features on Hadoop and Spark and make search results faster, thereby improving the speed of query results and enhancing the customer experience. You'll then get up to speed with performing analytics by building a metrics pipeline, defining queries, and using Kibana for intuitive visualizations that help provide decision-makers with better insights. The book will later guide you through using Logstash with examples to collect, parse, and enrich logs before indexing them in Elasticsearch. By the end of this book, you will have comprehensive knowledge of advanced topics such as Apache Spark support, machine learning using Elasticsearch and scikit-learn, and real-time analytics, along with the expertise you need to increase business productivity, perform analytics, and get the very best out of Elasticsearch. What you will learn

Pre-process documents before indexing in ingest pipelines
Learn how to model your data in the real world
Get to grips with using Elasticsearch for exploratory data analysis
Understand how to build analytics and RESTful services
Use Kibana, Logstash, and Beats for dashboard applications
Get up to speed with Spark and Elasticsearch for real-time analytics
Explore the basics of Spring Data Elasticsearch, and understand how to index, search, and query in a Spring application
Who this book is for
This book is for Elasticsearch developers and data engineers who want to take their basic knowledge of Elasticsearch to the next level and use it to build enterprise-grade distributed search applications. Prior experience of working with Elasticsearch will be useful to get the most out of this book.

Apache Solr - Dikshant Shahi 2015-12-26

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.

Enterprise Lucene and Solr - Lajos Moczar
2013-12-25

Increasingly, large enterprises are discovering that Apache Lucene and Apache Solr can help them dramatically reduce the costs of their most challenging indexing and search applications. Now, for the first time, there's a hands-on guide to using these technologies in enterprise environments. Packed with real-world examples and new best practices, *Enterprise Lucene and Solr* goes far beyond simply "getting started," to offer deep practical insights on planning, developing, and deploying highly-efficient solutions. This is practical, hands-on book. It is designed for people using Lucene and Solr in real-world, advanced applications. It covers topics such as enterprise search, spatial search, Machine Learning (specifically, text classification) and Big Data search. There is also a whole section devoted to what the author calls "productionalizing" Lucene and Solr, which will show you how to intelligently develop, deploy and maintain production installations. This book is divided into four parts. Part 1 covers the basics of Lucene and includes a chapter covering the basics, and then two more covering indexing and querying in depth, with hands-on examples in between. Part II provides a similar foundation for Solr. Part III is where the fun begins, because we get to build some interesting applications that demonstrate the power of these technologies. And finally, Part IV, covers the bit that many people forget about: how to use these technologies in production.

Apache Solr High Performance - Surendra Mohan
2014-03-25

This book is an easy-to-follow guide, full of

hands-on, real-world examples. Each topic is explained and demonstrated in a specific and user-friendly flow, from search optimization using Solr to Deployment of Zookeeper applications. This book is ideal for Apache Solr developers and want to learn different techniques to optimize Solr performance with utmost efficiency, along with effectively troubleshooting the problems that usually occur while trying to boost performance. Familiarity with search servers and database querying is expected.

Lucene in Action - Otis Gospodnetic
2010-07-08

When Lucene first hit the scene five years ago, it was nothing short of amazing. 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" feature into an indispensable part of most enterprise applications. Lucene now powers search in diverse companies including Akamai, Netflix, LinkedIn, Technorati, HotJobs, Epiphany, FedEx, Mayo Clinic, MIT, New Scientist Magazine, and many others. Some things remain the same, though. Lucene still delivers high-performance search features in a disarmingly easy-to-use API. Due to its vibrant and diverse open-source community of developers and users, Lucene is relentlessly improving, with evolutions to APIs, significant new features such as payloads, and a huge increase (as much as 8x) in indexing speed with Lucene 2.3. And with clear writing, reusable examples, and unmatched advice on best practices, *Lucene in Action, Second Edition* is still the definitive guide to developing 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.

Apache Solr 5.x Beginner's Guide - 2016

[Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications](#) - Gary Miner
2012-01-25
[Practical Text Mining and Statistical](#)

Analysis for Non-structured Text Data Applications brings together all the information, tools and methods a professional will need to efficiently use text mining applications and statistical analysis. Winner of a 2012 PROSE Award in Computing and Information Sciences from the Association of American Publishers, this book presents a comprehensive how-to-reference that shows the user how to conduct text mining and statistically analyze results. In addition to providing an in-depth examination of core text mining and link detection tools, methods and operations, the book examines advanced preprocessing techniques, knowledge representation considerations, and visualization approaches. Finally, the book explores current real-world, mission-critical applications of text mining and link detection using real world example tutorials in such varied fields as corporate, finance, business intelligence, genomics research, and counterterrorism activities. The world contains an unimaginably vast amount of digital information which is getting ever vaster ever more rapidly. This makes it possible to do many things that previously could not be done: spot business trends, prevent diseases, combat crime and so on. Managed well, the textual data can be used to unlock new sources of economic value, provide fresh insights into science and hold governments to account. As the Internet expands and our natural capacity to process the unstructured text that it contains diminishes, the value of text mining for information retrieval and search will increase dramatically. Extensive case studies, most in a tutorial format, allow the reader to 'click through' the example using a software program, thus learning to conduct text mining analyses in the most rapid manner of learning possible. Numerous examples, tutorials, power points and datasets available via companion website on Elsevierdirect.com Glossary of text mining terms provided in the appendix

Hadoop Security - Ben Spivey 2015-06-29
As more corporations turn to Hadoop to store and process their most valuable data,

the risk of a potential breach of those systems increases exponentially. This practical book not only shows Hadoop administrators and security architects how to protect Hadoop data from unauthorized access, it also shows how to limit the ability of an attacker to corrupt or modify data in the event of a security breach. Authors Ben Spivey and Joey Echeverria provide in-depth information about the security features available in Hadoop, and organize them according to common computer security concepts. You'll also get real-world examples that demonstrate how you can apply these concepts to your use cases. Understand the challenges of securing distributed systems, particularly Hadoop Use best practices for preparing Hadoop cluster hardware as securely as possible Get an overview of the Kerberos network authentication protocol Delve into authorization and accounting principles as they apply to Hadoop Learn how to use mechanisms to protect data in a Hadoop cluster, both in transit and at rest Integrate Hadoop data ingest into enterprise-wide security architecture Ensure that security architecture reaches all the way to end-user access

Intelligent Systems in Big Data, Semantic Web and Machine Learning - Noredine Gherabi 2021-05-28

This book describes important methodologies, tools and techniques from the fields of artificial intelligence, basically those which are based on relevant conceptual and formal development. The coverage is wide, ranging from machine learning to the use of data on the Semantic Web, with many new topics. The contributions are concerned with machine learning, big data, data processing in medicine, similarity processing in ontologies, semantic image analysis, as well as many applications including the use of machine learning techniques for cloud security, artificial intelligence techniques for detecting COVID-19, the Internet of things, etc. The book is meant to be a very important and useful source of information for researchers and doctoral students in

data analysis, Semantic Web, big data, machine learning, computer engineering and related disciplines, as well as for postgraduate students who want to integrate the doctoral cycle.

Natural Language Processing Recipes - Akshay Kulkarni 2019-01-29

Implement natural language processing applications with Python using a problem-solution approach. This book has numerous coding exercises that will help you to quickly deploy natural language processing techniques, such as text classification, parts of speech identification, topic modeling, text summarization, text generation, entity extraction, and sentiment analysis. *Natural Language Processing Recipes* starts by offering solutions for cleaning and preprocessing text data and ways to analyze it with advanced algorithms. You'll see practical applications of the semantic as well as syntactic analysis of text, as well as complex natural language processing approaches that involve text normalization, advanced preprocessing, POS tagging, and sentiment analysis. You will also learn various applications of machine learning and deep learning in natural language processing. By using the recipes in this book, you will have a toolbox of solutions to apply to your own projects in the real world, making your development time quicker and more efficient. *What You Will Learn* Apply NLP techniques using Python libraries such as NLTK, TextBlob, spaCy, Stanford CoreNLP, and many more. Implement the concepts of information retrieval, text summarization, sentiment analysis, and other advanced natural language processing techniques. Identify machine learning and deep learning techniques for natural language processing and natural language generation problems. *Who This Book Is For* Data scientists who want to refresh and learn various concepts of natural language processing through coding exercises.

Practical Hadoop Ecosystem - Deepak Vohra 2016-09-30

Learn how to use the Apache Hadoop projects, including MapReduce, HDFS,

Apache Hive, Apache HBase, Apache Kafka, Apache Mahout, and Apache Solr. From setting up the environment to running sample applications each chapter in this book is a practical tutorial on using an Apache Hadoop ecosystem project. While several books on Apache Hadoop are available, most are based on the main projects, MapReduce and HDFS, and none discusses the other Apache Hadoop ecosystem projects and how they all work together as a cohesive big data development platform. *What You Will Learn*: Set up the environment in Linux for Hadoop projects using Cloudera Hadoop Distribution CDH 5 Run a MapReduce job Store data with Apache Hive, and Apache HBase Index data in HDFS with Apache Solr Develop a Kafka messaging system Stream Logs to HDFS with Apache Flume Transfer data from MySQL database to Hive, HDFS, and HBase with Sqoop Create a Hive table over Apache Solr Develop a Mahout User Recommender System *Who This Book Is For*: Apache Hadoop developers. Pre-requisite knowledge of Linux and some knowledge of Hadoop is required.

Data Intensive Computing Applications for Big Data - M. Mittal 2018-01-31

The book 'Data Intensive Computing Applications for Big Data' discusses the technical concepts of big data, data intensive computing through machine learning, soft computing and parallel computing paradigms. It brings together researchers to report their latest results or progress in the development of the above mentioned areas. Since there are few books on this specific subject, the editors aim to provide a common platform for researchers working in this area to exhibit their novel findings. The book is intended as a reference work for advanced undergraduates and graduate students, as well as multidisciplinary, interdisciplinary and transdisciplinary research workers and scientists on the subjects of big data and cloud/parallel and distributed computing, and explains didactically many of the core concepts of these approaches for practical

applications. It is organized into 24 chapters providing a comprehensive overview of big data analysis using parallel computing and addresses the complete data science workflow in the cloud, as well as dealing with privacy issues and the challenges faced in a data-intensive cloud computing environment. The book explores both fundamental and high-level concepts, and will serve as a manual for those in the industry, while also helping beginners to understand the basic and advanced aspects of big data and cloud computing.

Open Source Intelligence Methods and Tools - Nihad A. Hassan 2018-06-30

Apply Open Source Intelligence (OSINT) techniques, methods, and tools to acquire information from publicly available online sources to support your intelligence analysis. Use the harvested data in different scenarios such as financial, crime, and terrorism investigations as well as performing business competition analysis and acquiring intelligence about individuals and other entities. This book will also improve your skills to acquire information online from both the regular Internet as well as the hidden web through its two sub-layers: the deep web and the dark web. The author includes many OSINT resources that can be used by intelligence agencies as well as by enterprises to monitor trends on a global level, identify risks, and gather competitor intelligence so more effective decisions can be made. You will discover techniques, methods, and tools that are equally used by hackers and penetration testers to gather intelligence about a specific target online. And you will be aware of how OSINT resources can be used in conducting social engineering attacks. Open Source Intelligence Methods and Tools takes a practical approach and lists hundreds of OSINT resources that can be used to gather intelligence from online public sources. The book also covers how to anonymize your digital identity online so you can conduct your searching activities without revealing your identity. What You'll Learn Identify intelligence needs and leverage a broad range of tools and sources

to improve data collection, analysis, and decision making in your organization Use OSINT resources to protect individuals and enterprises by discovering data that is online, exposed, and sensitive and hide the data before it is revealed by outside attackers Gather corporate intelligence about business competitors and predict future market directions Conduct advanced searches to gather intelligence from social media sites such as Facebook and Twitter Understand the different layers that make up the Internet and how to search within the invisible web which contains both the deep and the dark webs Who This Book Is For Penetration testers, digital forensics investigators, intelligence services, military, law enforcement, UN agencies, and for-profit/non-profit enterprises

Scaling Big Data with Hadoop and Solr - Second Edition - Hrishikesh Vijay Karambelkar 2015-04-30

Apache Kafka 1.0 Cookbook - Raúl Estrada 2017-12-22

Simplify real-time data processing by leveraging the power of Apache Kafka 1.0 Key Features Use Kafka 1.0 features such as Confluent platforms and Kafka streams to build efficient streaming data applications to handle and process your data Integrate Kafka with other Big Data tools such as Apache Hadoop, Apache Spark, and more Hands-on recipes to help you design, operate, maintain, and secure your Apache Kafka cluster with ease Book Description Apache Kafka provides a unified, high-throughput, low-latency platform to handle real-time data feeds. This book will show you how to use Kafka efficiently, and contains practical solutions to the common problems that developers and administrators usually face while working with it. This practical guide contains easy-to-follow recipes to help you set up, configure, and use Apache Kafka in the best possible manner. You will use Apache Kafka Consumers and Producers to build effective real-time streaming applications. The book covers the recently released Kafka version 1.0, the Confluent

Platform and Kafka Streams. The programming aspect covered in the book will teach you how to perform important tasks such as message validation, enrichment and composition. Recipes focusing on optimizing the performance of your Kafka cluster, and integrate Kafka with a variety of third-party tools such as Apache Hadoop, Apache Spark, and Elasticsearch will help ease your day to day collaboration with Kafka greatly. Finally, we cover tasks related to monitoring and securing your Apache Kafka cluster using tools such as Ganglia and Graphite. If you're looking to become the go-to person in your organization when it comes to working with Apache Kafka, this book is the only resource you need to have. What you will learn -Install and configure Apache

Kafka 1.0 to get optimal performance - Create and configure Kafka Producers and Consumers -Operate your Kafka clusters efficiently by implementing the mirroring technique -Work with the new Confluent platform and Kafka streams, and achieve high availability with Kafka -Monitor Kafka using tools such as Graphite and Ganglia - Integrate Kafka with third-party tools such as Elasticsearch, Logstash, Apache Hadoop, Apache Spark, and more Who this book is for This book is for developers and Kafka administrators who are looking for quick, practical solutions to problems encountered while operating, managing or monitoring Apache Kafka. If you are a developer, some knowledge of Scala or Java will help, while for administrators, some working knowledge of Kafka will be useful.