Instant Apache Solr For Indexing Data How To

LIZETH SANTOS

Solr Cookbook - Third Edition Packt Publishing Ltd
A fast paced guide that will help you learn about Apache Hadoop 3 and its ecosystem Key Features Set up, configure and get started with Hadoop to get useful insights from large data sets Work with the different components of Hadoop such as MapReduce, HDFS and YARN Learn about the new features introduced in Hadoop 3 Book Description Apache Hadoop is a widely used distributed data platform. It enables large datasets to be efficiently processed instead of using one large computer to store and process the data. This book will get you started with the Hadoop ecosystem, and introduce you to the main technical topics, including MapReduce, YARN, and HDFS. The book begins with an overview of big data and Apache Hadoop. Then, you will set up a pseudo Hadoop development environment and a multi-node enterprise Hadoop cluster. You will see how the parallel programming paradigm, such as MapReduce, can solve many complex data processing problems. The book also covers the important aspects of the big data software development lifecycle, including quality assurance and control, performance, administration, and monitoring. You will then learn about the Hadoop ecosystem, and tools such as Kafka, Sqoop, Flume, Pig, Hive, and HBase. Finally, you will look at advanced topics, including real time streaming using Apache Storm, and data analytics using Apache Spark. By the end of the book, you will be well versed with different configurations of the Hadoop 3 cluster. What you will learn Store and analyze data at scale using HDFS, MapReduce and YARN Install and configure Hadoop 3 in different modes Use Yarn effectively to run different applications on Hadoop based platform Understand and monitor how Hadoop cluster is managed Consume streaming data using Storm, and then analyze it using Spark Explore Apache Hadoop ecosystem components, such as Flume, Sqoop, HBase, Hive, and Kafka Who this book is for Aspiring Big Data professionals who want to learn the essentials of Hadoop 3 will find this book to be useful. Existing Hadoop users who want to get up to speed with the new features introduced in Hadoop 3 will also benefit from this book. Having knowledge of Java programming will be an added advantage.
Instant Apache Solr for Indexing Data How-to Packt Publishing Ltd
A step-by-step tutorial on implementing Liferay- based portals to learn performance best practices.The book is good for Liferay portal developers and architects who want to learn performance best practices for implementing Liferay- based solutions. It is assumed that you have a working knowledge of the Liferay portal.
Cassandra: The Definitive Guide Packt Publishing Ltd
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.
Tika in Action Springer Science & Business Media
Summary Making Sense of NoSQL clearly and concisely explains the concepts, features, benefits, potential, and limitations of NoSQL technologies. Using examples and use cases, illustrations, and plain, jargon-free writing, this guide shows how you can effectively assemble a NoSQL solution to replace or augment the traditional RDBMS you have now. About this Book If you want to understand and perhaps start using the new data storage and analysis technologies that go beyond the SQL database model, this book is for you. Written in plain language suitable for technical managers and developers, and using many examples, use cases, and illustrations, this book explains the concepts, features, benefits, potential, and limitations of NoSQL. Making Sense of NoSQL starts by comparing familiar database concepts to the new NoSQL patterns that augment or replace them. Then, you'll explore case studies on big data, search, reliability, and business agility that apply these new patterns to today's business problems. You'll see how NoSQL systems can leverage the resources of modern cloud computing and multiple-CPU data centers. The final chaptersshow you how to choose the right NoSQL technologies for your own needs. Managers and developers will welcome this lucid overview of the potential and capabilities of NoSQL technologies. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. What's Inside NoSQL data architecture patterns NoSQL for big data Search, high availability, and security Choosing an architecture About the Authors Dan McCreary and Ann Kelly lead an independent training and consultancy firm focused on NoSQL solutions and are cofounders of the NoSQL Now! Conference. Table of Contents PART 1 INTRODUCTION NoSQL: It's about making intelligent choices NoSQL concepts PART 2 DATABASE PATTERNS Foundational data architecture patterns NoSQL data architecture patterns Native XML databases PART 3 NOSQL SOLUTIONS Using NoSQL to manage big data Finding information with NoSQL search Building high-availability solutions with NoSQL Increasing agility with NoSQL PART 4 ADVANCED TOPICS NoSQL and functional programming Security: protecting data in your NoSQL systems Selecting the right NoSQL solution
Solr 1.4 Enterprise Search Server Packt Publishing Ltd
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
Finding Source Code on the Web for Remix and Reuse "O'Reilly Media, Inc."
A comprehensive guide to mastering the most advanced Hadoop 3 concepts Key Features Get to grips with the newly introduced features and capabilities of Hadoop 3 Crunch and process data using MapReduce, YARN, and a host of tools within the Hadoop ecosystem Sharpen your Hadoop skills with real-world case studies and code Book Description Apache Hadoop is one of the most popular big data solutions for distributed storage and for processing large chunks of data. With Hadoop 3, Apache promises to provide a high-performance, more fault-tolerant, and highly efficient big data processing platform, with a focus on improved scalability and increased efficiency. With this guide, you'll understand advanced concepts of the Hadoop ecosystem tool. You'll learn how Hadoop works internally, study advanced concepts of different ecosystem tools, discover solutions to real-world

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

Design Patterns for Cloud Native Applications John Wiley & Sons

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.

Apache Solr Search Patterns Cambridge University Press

Over 100 practical recipes to make Apache Solr faster, more reliable and return better results.

Alfresco for Administrators Packt Pub Limited

This book is a step-by-step tutorial that will enable you to leverage the flexible search functionality of Apache Solr together with the Big Data power of Apache Hadoop.Scaling Big Data with Hadoop and Solr provides guidance to developers who wish to build high-speed enterprise search platforms using Hadoop and Solr. This book is primarily aimed at Java programmers who wish to extend the Hadoop platform to make it run as an enterprise search without any prior knowledge of Apache Hadoop and Solr.

Spring Data Instant Apache Solr for Indexing Data How-toFilled with practical, step-by-step instructions and clear explanations for the most important and useful tasks. This book is written in a friendly, practical manner with recipes covering important indexing techniques and methods using Apache Solr.This book is for developers who want to dive deeper into Solr. Regardless of whether you are just starting with Solr or have already built your first collection by copying and modifying examples, this book will take you through the complicated steps of indexing your data with Solr.Apache Solr for Indexing Data

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

How to Find, Organize, and Manipulate It "O'Reilly Media, Inc."

You can choose several data access frameworks when building Java enterprise applications that work with relational databases. But what about big data? This hands-on introduction shows you how Spring Data makes it relatively easy to build applications across a wide range of new data access technologies such as NoSQL and Hadoop. Through several sample projects, you'll learn how Spring Data provides a consistent programming model that retains NoSQL-specific features and capabilities, and helps you develop Hadoop applications across a wide range of use-cases such as data analysis, event stream processing, and workflow. You'll also discover the features Spring Data adds to Spring's existing JPA and JDBC support for writing RDBMS-based data access layers. Learn about Spring's template helper classes to simplify the use of database-specific functionality Explore Spring Data's repository abstraction and advanced query functionality Use Spring Data with Redis (key/value store), HBase (column-family), MongoDB (document database), and Neo4j (graph database) Discover the GemFire distributed data grid solution Export Spring Data JPA-managed entities to the Web as RESTful web services Simplify the development of HBase applications, using a lightweight object-mapping framework Build example big-data pipelines with Spring Batch and Spring Integration

Mastering Apache Solr 7. X "O'Reilly Media, Inc."

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

Introduction to Information Retrieval Packt Publishing Ltd

Apache Solr: Classic Edition. There has never been a Apache Solr Guide like this. It contains 31 answers, much more than you can imagine; comprehensive answers and extensive details and references, with insights that have never before been offered in print. Get the information you need--fast! This all-embracing guide offers a thorough view of key knowledge and detailed insight. This Guide introduces what you want to know about Apache Solr. A quick look inside of some of the subjects covered: LucidWorks, Spring Roo - Motivation and History, Universally unique identifier - Implementations, LucidWorks - Awards, Apache Cassandra - Integration with other tools, UUID - Implementations, Giant Bomb - Development, GoPivotal - Analytics, CiteSeer - CiteSeerX, Spring Roo - Standards and Technology Compatibility, Norconex HTTP Collector - Architecture, Solr, Apache Solr, Riak - Main Features, Open Search Server - Competitors, Riak - History, Homeland Security Digital Library - Technology, ColdFusion - Adobe ColdFusion 10, ColdFusion - Main features, Full text search Free and open source software, Spring Roo - User Interface, Faceted search - Projects, Full-text search - Free and open source software, CiteSeerX - CiteSeerX, Lucene.net - Lucene-based projects, List of enterprise search vendors - Free and open source enterprise search software, Apache Cassandra Integration with other tools, Comic Vine - Development, Apache Solr - History, List of information retrieval libraries - Libraries for searching and indexing, and much more...

Apache Solr for Indexing Data Packt Publishing Ltd

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.

Apache Solr 4 Cookbook Emereo Publishing

Summary Redis in Action introduces Redis and walks you through examples that demonstrate how to use it effectively. You'll begin by getting Redis set up properly and then exploring the key-value model. Then, you'll dive into real use cases including simple caching, distributed ad targeting, and more. You'll learn how to scale Redis from small jobs to massive datasets. Experienced developers will appreciate chapters on clustering and internal scripting to make Redis easier to use. About the Technology When you need near-real-time access to a fast-moving data stream, key-value stores like Redis are the way to go. Redis expands on the key-value pattern by accepting a wide variety of data types, including hashes, strings, lists, and other structures. It provides lightning-fast operations on in-memory datasets, and also makes it easy to persist to disk on the fly. Plus, it's free and open source. About this book Redis in Action introduces Redis and the key-value model. You'll quickly dive into real use cases including simple caching, distributed ad targeting, and more. You'll learn how to scale Redis from small jobs to massive datasets and discover how to integrate with traditional RDBMS or other NoSQL stores. Experienced developers will appreciate the in-depth chapters on clustering and internal scripting. Written for developers familiar with database concepts. No prior exposure to NoSQL database concepts nor to Redis itself is required. Appropriate for systems administrators comfortable with programming. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. What's Inside Redis from the ground up Preprocessing real-time data Managing in-memory datasets Pub/sub and configuration Persisting to disk About the Author Dr. Josiah L. Carlson is a seasoned database professional and an active contributor to the Redis community. Table of Contents PART 1 GETTING STARTED Getting to know Redis Anatomy of a Redis web application PART 2 CORE CONCEPTS Commands in Redis Keeping data safe and ensuring performance Using Redis for application support Application components in Redis Search-based applications Building a simple social network PART 3 NEXT STEPS Reducing memory use Scaling Redis Scripting Redis with Lua

Making Sense of NoSQL Packt Publishing Ltd

A fast-paced administrator's guide to Alfresco from the administration, managing, and high-level design perspectives About This Book Understand system capabilities in order to make informed and appropriate decisions about its administration Manage users, groups, email, file systems, and

transformer availability using Alfresco Use Alfresco to capture and efficiently manage information about repositories, servers, and statistics Who This Book Is For The target audience would be users with a basic knowledge of Content Management System, and also users who want to understand Alfresco from the administration and high-level design perspectives. What You Will Learn Understand Alfresco's architecture and important building blocks Learn to install Alfresco on various application servers such as Tomcat , JBoss, and WebLogic. Become familiar with various configurations in Alfresco such as databases, filesystems, email, and audits Administrate Alfresco using the Explorer Admin Console, Share Admin Console, and Workflow Admin Console Understand how to integrate LDAP and Active Directory with Alfresco for centralized user management Learn how Alfresco environments can be clustered for high availability Fully understand how Alfresco stores content and easily retrieve any information from Alfresco Monitor and manage Alfresco systems in production In Detail Alfresco is an open source Enterprise Content Management (ECM) system for Windows and Linux-like operating systems. The year-on-year growth of business connections, contacts, and communications is expanding enterprise boundaries more than ever before. Alfresco enables organizations to collaborate more effectively, improve business process efficiency, and ensure information governance. The basic purpose of Alfresco is to help users to capture and manage information in a better way. It helps you capture, organize, and share binary files. This book will cover the basic building blocks of an Alfresco system, how the components fit together, and the information required to build a system architecture. This book will also focus on security aspects of Alfresco. such as authentication, troubleshooting, managing permissions, and so on. It will also focus on managing content and storage, indexing and searches, setting up clustering for high availability, and so forth. Style and approach A step-by-step guide to understanding the Alfresco system and making informed and appropriate decisions about administration.

A Distributed Real-Time Search and Analytics Engine Simon and Schuster

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.

Solr in Action Packt Pub Limited

Instant Apache Solr for Indexing Data How-to

Mastering Hadoop 3 Simon and Schuster

With the immense cost savings and scalability the cloud provides, the rationale for building cloud native applications is no longer in question. The real issue is how. With this practical guide, developers will learn about the most commonly used design patterns for building cloud native applications using APIs, data, events, and streams in both greenfield and brownfield development. You'll learn how to incrementally design, develop, and deploy large and effective cloud native applications that you can manage and maintain at scale with minimal cost, time, and effort. Authors Kasun Indrasiri and Sriskandarajah Suhothayan highlight use cases that effectively demonstrate the challenges you might encounter at each step. Learn the fundamentals of cloud native applications Explore key cloud native communication, connectivity, and composition patterns Learn decentralized data management techniques Use event-driven architecture to build distributed and scalable cloud native applications Explore the most commonly used patterns for API management and consumption Examine some of the tools and technologies you'll need for building cloud native systems

Mastering Elasticsearch - Second Edition Simon and Schuster

Written in a friendly, example-driven format, the book includes plenty of step-by-step instructions and examples that are designed to help you get started with Apache Solr.This book is an entry level text into the wonderful world of Apache Solr. The book will center around a couple of simple projects such as setting up Solr and all the stuff that comes with customizing the Solr schema and configuration. This book is for developers looking to start using Apache Solr who are stuck or intimidated by the difficulty of setting it up and using it.For anyone wanting to embed a search engine in their site to help users navigate around the mammoth data available this book is an ideal starting point. Moreover, if you are a data architect or a project manager and want to make some key design decisions, you will find that every example included in the book contains ideas usable in real-world contexts.