Instant Apache Solr For Indexing Data How To

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 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.

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

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

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...

If you are a developer who implements ElasticSearch in your web applications and want to sharpen your understanding of the core elements and applications, this is the book for you. It is assumed that you've got working knowledge of JSON and, if you want to extend ElasticSearch, of Java and related technologies.

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.

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.

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.
Enterprise and web applications require full-featured, "Google-quality" searchcapabilities, but such features are notoriously difficult to implement and maintain.Hibernate Search builds on the Lucene feature set and offers an easyto-implement interface that integrates seamlessly with Hibernate-the leadingdata persistence solution for Java applications. Hibernate Search in Action introduces both the principles of enterprise searchand the implementation details a Java developer will need to use HibernateSearch effectively. This book blends the insights of the Hibernate Search leaddeveloper with the practical techniques required to index and manipulate data,assemble and execute search queries, and create smart filters for better searchresults. Along the way, the reader masters performance-boosting concepts likeusing Hibernate Search in a clustered environment and integrating with thefeatures already in your applications. This book assumes you're a competent Java developer with some experienceusing Hibernate and 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.
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.
ElasticSearch is an open source search server built on Apache Lucene. It was built to provide a scalable search solution with built-in support for near real-time search and multi-tenancy.Jumping into the world of ElasticSearch by setting up your own custom cluster, this book will show you how to create a fast, scalable, and flexible search solution. By learning the ins-and-outs of data indexing and analysis, "ElasticSearch Server" will start you on your journey to mastering the powerful capabilities of ElasticSearch. With practical chapters covering how to search data, extend your search, and go deep into cluster administration and search analysis, this book is perfect for those new and experienced with search servers.In "ElasticSearch Server" you will learn how to revolutionize your website or application with faster, more accurate, and flexible search functionality. Starting with chapters on setting up your own ElasticSearch cluster and searching and extending your search parameters you will quickly be able to create a fast, scalable, and completely custom search solution.Building on your knowledge further you will learn about ElasticSearch's query API and become confident using powerful filtering and faceting capabilities. You will develop practical knowledge on how to make use of ElasticSearch's near real-time capabilities and support for multi-tenancy.Your journey then concludes with chapters that help you monitor and tune your ElasticSearch cluster as well as advanced topics such as shard allocation, gateway configuration, and the discovery module.
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
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.

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

A hands-on guide to leveraging NoSQL databases NoSQL databases are an efficient and powerful tool for storing and manipulating vast quantities of data. Most NoSQL databases scale well as data grows. In addition, they are often malleable and flexible enough to accommodate semi-structured and sparse data sets. This comprehensive hands-on guide presents fundamental concepts and practical solutions for getting you ready to use NoSQL databases. Expert author Shashank Tiwari begins with a helpful introduction on the subject of NoSQL, explains its characteristics and typical uses, and looks at where it fits in the application stack. Unique insights help you choose which NoSQL solutions are best for solving your specific data storage needs. Professional NoSQL: Demystifies the concepts that relate to NoSQL databases, including column-family oriented stores, key/value databases, and document databases. Delves into installing and configuring a number of NoSQL products and the Hadoop family of products. Explains ways of storing, accessing, and querying data in NoSQL databases through examples that use MongoDB, HBase, Cassandra, Redis, CouchDB, Google App Engine Datastore and more. Looks at architecture and internals. Provides guidelines for optimal usage, performance tuning, and scalable configurations. Presents a number of tools and utilities relating to NoSQL, distributed platforms, and scalable processing, including Hive, Pig, RRDtool, Nagios, and more.

If you are a system or application developer interested in learning how to solve practical problems using the Hadoop framework, then this book is ideal for you. This book is also meant for Hadoop professionals who want to find solutions to the different challenges they come across in their Hadoop projects.

This book is for Elasticsearch users who want to extend their knowledge and develop new skills. Prior knowledge of the Query DSL and data indexing is expected.

Learn the art of efficient web scraping and crawling with Python About This Book Extract data from any source to perform real time analytics. Full of techniques and examples to help you crawl websites and extract data within hours. A hands-on guide to web scraping and crawling with real-life problems and solutions Who This Book Is For If you are a software developer, data scientist, NLP or machine-learning enthusiast or just need to migrate your company's wiki from a legacy platform, then this book is for you. It is perfect for someone , who needs instant access to large amounts of semi-structured data effortlessly. What You Will Learn Understand HTML pages and write XPath to extract the data you need Write Scrapy spiders with simple Python and do web crawls Push your data into any database, search engine or analytics system Configure your spider to download files, images and use proxies Create efficient pipelines that shape data in precisely the form you want Use Twisted Asynchronous API to process hundreds of items concurrently Make your crawler super-fast by learning how to tune Scrapy's performance Perform large scale distributed crawls with scrapyd and scrapinghub In Detail This book covers the long awaited Scrapy v 1.0 that empowers you to extract useful data from virtually any source with very little effort. It starts off by explaining the fundamentals of Scrapy framework, followed by a thorough description of how to extract data from any source, clean it up, shape it as per your requirement using Python and 3rd party APIs. Next you will be familiarised with the process of storing the scrapped data in databases as well as search engines and performing real time analytics on them with Spark Streaming. By the end of this book, you will perfect the art of scarping data for your applications with ease Style and approach It is a hands on guide, with first few chapters written as a tutorial, aiming to motivate you and get you started quickly. As the book progresses, more advanced features are explained with real world examples that can be reffered while developing your own web applications.

Deliver end-to-end real-time distributed data processing solutions by leveraging the power of Elastic Stack 6.0 Key Features - Get to grips with the new features introduced in Elastic Stack 6.0 - Get valuable insights from your data by working with the different components of the Elastic stack such as Elasticsearch, Logstash, Kibana, X-Pack, and Beats - Includes handy tips and techniques to build, deploy and manage your Elastic applications efficiently on-premise or on the cloud Book Description The Elastic Stack is a powerful combination of tools for distributed search, analytics, logging, and visualization of data from medium to massive data sets. The newly released Elastic Stack 6.0 brings new features and capabilities that empower users to find unique, actionable insights through these techniques. This book will give you a fundamental understanding of what the stack is all about, and how to use it efficiently to build powerful real-time data processing applications. After a quick overview of the newly introduced features in Elastic Stack 6.0, you'll learn how to set up the stack by installing the tools, and see their basic configurations. Then it shows you how to use Elasticsearch for distributed searching and analytics, along with Logstash for logging, and Kibana for data visualization. It also demonstrates the creation of custom plugins using Kibana and Beats. You'll find out about Elastic X-Pack, a useful extension for effective security and monitoring. We also provide useful tips on how to use the Elastic Cloud and deploy the Elastic Stack in production environments. On completing this book, you'll have a solid foundational knowledge of the basic Elastic Stack functionalities. You'll also have a good understanding of the role of each component in the stack to solve different data processing problems. What you will learn - Familiarize yourself with the different components of the Elastic Stack - Get to know the new functionalities introduced in Elastic Stack 6.0 - Effectively build your data pipeline to get data from terabytes or petabytes of data into Elasticsearch and Logstash for searching and logging - Use Kibana to visualize data and tell data stories in real-time - Secure, monitor, and use the alerting and reporting capabilities of Elastic Stack - Take your Elastic application to an on-premise or cloud-based production environment Who this book is for This book is for data professionals who want to get amazing insights and business metrics from their data sources. If you want to get a fundamental understanding of the Elastic Stack for distributed, real-time processing of data, this book will help you. A fundamental knowledge of JSON would be useful, but is not mandatory. No previous experience with the Elastic Stack is required.

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

What could you do with data if scalability wasn't a problem? With this hands-on guide, you'll learn how Apache Cassandra handles hundreds of terabytes of data while remaining highly available across multiple data centers -- capabilities that have attracted Facebook, Twitter, and other data-intensive companies. Cassandra: The Definitive Guide provides the technical details and practical examples you need to assess this database management system and put it to work in a production environment. Author Eben Hewitt demonstrates the advantages of Cassandra's nonrelational design, and pays special attention to data modeling. If you're a developer, DBA, application architect, or manager looking to solve a database scaling issue or future-proof your application, this guide shows you how to harness Cassandra's speed and flexibility. Understand the tenets of Cassandra's column-oriented structure Learn how to write, update, and read Cassandra data Discover how to add or remove nodes from the cluster as your application requires Examine a working application that translates from a relational model to Cassandra's data model Use examples for writing clients in Java, Python, and C# Use the JMX interface to monitor a cluster's usage, memory patterns, and more Tune memory settings, data storage, and caching for better performance

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.

This open access book covers all facets of entity-oriented search—where “search” can be interpreted in the broadest sense of information access—from a unified point of view, and provides a coherent and comprehensive overview of the state of the art. It represents the first synthesis of research in this broad and rapidly developing area. Selected topics are discussed in-depth, the goal being to establish fundamental techniques and methods as a basis for future research and development. Additional topics are treated at a survey level only, containing numerous pointers to the relevant literature. A roadmap for future research, based on open issues and challenges identified along the way, rounds out the book. The book is divided into three main parts, sandwiched between introductory and concluding chapters. The first two chapters introduce readers to the basic concepts, provide an overview of entity-oriented search tasks, and present the various types and sources of data that will be used throughout the book. Part I deals with the core task of entity ranking: given a textual query, possibly enriched with additional elements or structural hints, return a ranked list of entities. This core task is examined in a number of different variants, using both structured and unstructured data collections, and numerous query formulations. In turn, Part II is devoted to the role of entities in bridging unstructured and structured data. Part III explores how entities can enable search engines to understand the concepts, meaning, and intent behind the query that the user enters into the search box, and how they can provide rich and focused responses (as opposed to merely a list of documents)—a process known as semantic search. The final chapter concludes the book by discussing the limitations of current approaches, and suggesting directions for future research. Researchers and graduate students are the primary target audience of this book. A general background in information retrieval is sufficient to follow the material, including an understanding of basic probability and statistics concepts as well as a basic knowledge of machine learning concepts and supervised learning algorithms.

In this book readers will find technological discussions on the existing and emerging technologies across the different stages of the big data value chain. They will learn about legal aspects of big data, the social impact, and about education needs and requirements. And they will discover the business perspective and how big data technology can be exploited to deliver value within different sectors of the economy. The book is structured in four parts: Part I “The Big Data Opportunity” explores the value potential of big data with a particular focus on the European context. It also describes the legal, business and social dimensions that need to be addressed, and briefly introduces the European Commission’s BIG project. Part II “The Big Data Value Chain” details the complete big data lifecycle from a technical point of view, ranging from data acquisition, analysis, curation and storage, to data usage and exploitation. Next, Part III “Usage and Exploitation of Big Data” illustrates the value creation possibilities of big data applications in various sectors, including industry, healthcare, finance, energy, media and public services. Finally, Part IV “A Roadmap for Big Data Research” identifies and prioritizes the cross-sectorial requirements for big data research, and outlines the most urgent and challenging technological, economic, political and societal issues for big data in Europe. This compendium summarizes more than two years of work performed by a leading group of major European research centers and industries in the context of the BIG project. It brings together research findings, forecasts and estimates related to this challenging technological context that is becoming the major axis of the new digitally transformed business environment.

JSON is becoming the backbone for meaningful data interchange over the internet. This format is now supported by an entire ecosystem of standards, tools, and technologies for building truly elegant, useful, and efficient applications. With this hands-on guide, author and architect Tom Marrs shows you how to build enterprise-class applications and services by leveraging JSON tooling and message/document design. JSON at Work provides application architects and developers with guidelines, best practices, and use cases, along with lots of real-world examples and code samples. You’ll start with a comprehensive JSON overview, explore the JSON ecosystem, and then dive into JSON’s use in the enterprise. Get acquainted with JSON basics and learn how to model JSON data Learn how to use JSON with Node.js, Ruby on Rails, and Java Structure JSON documents with JSON Schema to design and test APIs Search the contents of JSON documents with JSON Search tools Convert JSON documents to other data formats with JSON Transform tools Compare JSON-based hypermedia formats, including HAL and jsonapi Leverage MongoDB to store and access JSON documents Use Apache Kafka to exchange JSON-based messages between services

Summary Relevant Search demystifies relevance work. Using Elasticsearch, it teaches you how to return engaging search results to your users, helping you understand and leverage the internals of Lucene-based search engines. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology Users are accustomed to and expect instant, relevant search results. To achieve this, you must master the search engine. Yet for many developers, relevance ranking is mysterious or confusing. About the Book Relevant Search demystifies the subject and shows you that a search engine is a programmable relevance framework. You'll learn how to apply Elasticsearch or Solr to your business's unique ranking problems. The book demonstrates how to program relevance and how to incorporate secondary data sources, taxonomies, text analytics, and personalization. In practice, a relevance framework requires softer skills as well, such as collaborating with stakeholders to discover the right relevance requirements for your business. By the end, you'll be able to achieve a virtuous cycle of provable, measurable relevance improvements over a search product's lifetime. What's Inside Techniques for debugging relevance? Applying search engine features to real problems? Using the user interface to guide searchers? A systematic approach to relevance? A business culture focused on improving search About the Reader For developers trying to build smarter search with Elasticsearch or Solr. About the Authors Doug Turnbull is lead relevance consultant at OpenSource Connections, where he frequently speaks and blogs. John Berryman is a data engineer at Eventbrite, where he specializes in recommendations and search. Foreword author, Trey Grainger, is a director of engineering at CareerBuilder and author of Solr in Action. Table of Contents The search relevance problem Search under the hood Debugging your first relevance problem Taming tokens Basic multifield search Term-centric search Shaping the relevance function Providing relevance feedback Designing a relevance-focused search application The relevance-centered enterprise Semantic and personalized search

Filled with practical, step-by-step instructions and clear explanations for the most important and useful tasks. Instant Apache Stanbol How-to will enable you to become an expert in content management with semantics at the core, with the help of practical recipes. Instant Apache Stanbol How-to is for Java developers who would like to extend Stanbol or would just like to use Stanbol without caring about its internals. A few recipes that show how to extend Stanbol require some familiarity with Java and JavaScript.

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.

Filled 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.

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

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.

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.

Lucene 4 Cookbook is a practical guide that shows you how to build a scalable search engine for your application, from an internal documentation search to a wide-scale web implementation with millions of records. Starting with helping you to successfully install Apache Lucene, it will guide you through creating your first search application.

Furthermore, the book walks you through analyzing your text and indexing your data to leverage the performance of your search application. As you progress through the chapters, you will learn to effectively search your indexes and successfully employ real-time searching. The chapters start off with simple concepts and build up to complex solutions that should help you on your way to becoming a search engine expert.

This book is part of Packt's Cookbook series; each chapter looks at a different aspect of working with Apache Solr. The recipes deal with common problems of working with Solr by using easy-to-understand, real-life examples. The book is not in any way a complete Apache Solr reference and you should see it as a helping hand when things get rough on your journey with Apache Solr. Developers who are working with Apache Solr and would like to know how to combat common problems will find this book of great use. Knowledge of Apache Lucene would be a bonus but is not required.

Get up to speed on the nuances of NoSQL databases and what they mean for your organization This easy to read guide to NoSQL databases provides the type of no-nonsense overview and analysis that you need to learn, including what NoSQL is and which database is right for you. Featuring specific evaluation criteria for NoSQL databases, along with a look into the pros and cons of the most popular options, NoSQL For Dummies provides the fastest and easiest way to dive into the details of this incredible technology. You'll gain an understanding of how to use NoSQL databases for mission-critical enterprise architectures and projects, and real-world examples reinforce the primary points to create an action-oriented resource for IT pros. If you're planning a big data project or platform, you probably already know you need to select a NoSQL database to complete your architecture. But with options flooding the market and updates and add-ons coming at a rapid pace, determining what you require now, and in the future, can be a tall task. This is where NoSQL For Dummies comes in! Learn the basic tenets of NoSQL databases and why they have come to the forefront as data has outpaced the capabilities of relational databases Discover major players among NoSQL databases, including Cassandra, MongoDB, MarkLogic, Neo4J, and others Get an in-depth look at the benefits and disadvantages of the wide variety of NoSQL database options Explore the needs of your organization as they relate to the capabilities of specific NoSQL databases Big data and Hadoop get all the attention, but when it comes down to it, NoSQL databases are the engines that power many big data analytics initiatives. With NoSQL For Dummies, you'll go beyond relational databases to ramp up your enterprise's data architecture in no time.

In recent years, searching for source code on the web has become increasingly common among professional software developers and is emerging as an area of academic research. This volume surveys past research and presents the state of the art in the area of "code retrieval on the web." This work is concerned with the algorithms, systems, and tools to allow programmers to search for source code on the web and the empirical studies of these inventions and practices. It is a label that we apply to a set of related research from software engineering, information retrieval, human-computer interaction, management, as well as commercial products. The division of code retrieval on the web into snippet remixing and component reuse is driven both by empirical data, and analysis of existing search engines and tools. Contributors include leading researchers from human-computer interaction, software engineering, programming languages, and management. "Finding Source Code on the Web for Remix and Reuse" consists of five parts. Part I is titled "Programmers and Practices," and consists of a retrospective chapter and two empirical studies on how programmers search the web for source code. Part II is titled "From Data Structures to Infrastructures," and covers the creation of ground-breaking search engines for code retrieval required ingenuity in the adaptation of existing technology and in the creation of new algorithms and data structures. Part III focuses on "Reuse: Components and Projects," which are reused with minimal modification. Part IV is on "Remix: Snippets and Answers," which examines how source code from the web can also be used as solutions to problems and answers to questions. The book concludes with Part V, "Looking Ahead," that looks at future programming and the legalities of software reuse and remix and the implications of current intellectual property law on the future of software development. The story, "Richie Boss: Private Investigator Manager," was selected as the winner of a crowdfunded short story contest."

This book is for intermediate Solr Developers who are willing to learn and implement Pro-level practices, techniques, and solutions. This edition will specifically appeal to developers who wish to quickly get to grips with the changes and new features of Apache Solr 5.

Let Hadoop For Dummies help harness the power of your data and rein in the information overload Big data has become big business, and companies and organizations of all sizes are struggling to find ways to retrieve valuable information from their massive data sets with becoming overwhelmed. Enter Hadoop and this easy-to-understand For Dummies guide. Hadoop For Dummies helps readers understand the value of big data, make a business case for using Hadoop, navigate the Hadoop ecosystem, and build and manage Hadoop applications and clusters. Explains the origins of Hadoop, its economic benefits, and its functionality and practical applications Helps you find your way around the Hadoop ecosystem, program MapReduce, utilize design patterns, and get your Hadoop cluster up and running quickly and easily Details how to use Hadoop applications for data mining, web analytics and personalization, large-scale text processing, data science, and problem-solving Shows you how to improve the value of your Hadoop cluster, maximize your investment in Hadoop, and avoid common pitfalls when building your Hadoop cluster From programmers challenged with building and maintaining affordable, scaleable data systems to administrators who must deal with huge volumes of information effectively and efficiently, this how-to has something to help you with Hadoop.