实战 Elasticsearch、Logstash、Kibana

——分布式大数据搜索与日志挖掘及可视化解决方案

高 凯 编著

清华大学出版社

实战 Elasticsearch、Logstash、Kibana

——分布式大数据搜索与日志挖掘及可视化解决方案

高　凯 编著

清华大学出版社
北京

内 容 简 介

对大数据的搜索与挖掘，在当今网络时代是很有必要的。本书提出的分布式大数据搜索与日志挖掘及可视化解决方案是基于 Elasticsearch、Logstash 和 Kibana 而形成的，它能有效应对海量大数据所带来的分布式存储与处理、全文检索、日志挖掘、可视化等的挑战。构建在全文检索开源软件 Lucene 之上的 Elasticsearch，不仅能对海量规模的数据完成分布式索引与检索，还能提供数据聚合分析；Logstash 能有效处理来源于各种不同数据源的日志信息；Kibana 能得出可视化分析结果。本书讲解有关 Elasticsearch、Logstash、Kibana 的使用，相关内容以模块化的方式进行组织，注重实战，强调实践，内容新颖，组织合理。

本书可为高校相关专业（如计算机科学与技术、软件工程、情报学、图书馆学、信息管理与信息系统）学生的学习和科研工作提供帮助，同时对于从事大数据搜索与挖掘、信息检索与智能处理技术的工程技术人员和希望了解网络信息检索与分析技术的爱好者也具有较高的参考价值。

图书在版编目(CIP)数据

实战 Elasticsearch、Logstash、Kibana——分布式大数据搜索与日志挖掘及可视化解决方案/高凯编著. --北京：清华大学出版社，2015

ISBN 978-7-302-39984-1

Ⅰ. ①实…　Ⅱ. ①高…　Ⅲ. ①互联网络-信息检索　Ⅳ. ①G354.4

中国版本图书馆 CIP 数据核字（2015）第 115651 号

责任编辑：焦　虹　李　晔
封面设计：傅瑞学
责任校对：徐俊伟
责任印制：杨　艳

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社　总　机：010-62770175　　邮　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课 件 下 载：http://www.tup.com.cn，010-62795954

印 装 者：北京鑫海金澳胶印有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：15.25　　**字　　数**：371 千字
版　　次：2015 年 6 月第 1 版　　**印　　次**：2015 年 6 月第 1 次印刷
印　　数：1～2000
定　　价：49.00 元

产品编号：062546-01

序言

Preface

云计算、智慧城市、移动互联网、大数据与物联网已经成为大数据时代的前瞻技术，实现了人、机器与实物的多维互联互通，监测数据、内容数据、社交数据、关系数据裂变式增长，大数据时代全方位地到来。大数据具有多（体量大）、快（生成速度快）、好（价值大）、省（高效）的特征，传统的信息搜索、数据挖掘与知识呈现理论技术难以满足当下多样化的需求。大数据的理念与理论已经成为了人所共知的科学常识，但是大数据搜索、挖掘与可视化等落地的工程实践尚有较大距离，也是当下的工程急需。

本书从分布式大数据搜索、日志挖掘与可视化三个角度出发，以非结构化文本信息、半结构化的日志数据为处理对象，进行宏观解决方案与微观方法技巧全面阐释。具体地说，如何利用在全文检索开源软件 Lucene 之上的 Elasticsearch 对大数据进行分布式计算与全文检索；如何利用 Logstash 对日志文件智能分析与处理；如何利用 Web 接口 Kibana 对日志进行高效的搜索、可视化、分析等各种操作是，是本书的论述重点。

从工程实践的角度掌握 ElasticSearch、Logstash、Kibana 的基本使用方法和技巧，很有必要。目前，国内专门针对 ElasticSearch、Logstash、Kibana 进行介绍的书很少，本书是目前国内较早的一本综合介绍 ELK 架构的编著，涉及范围广泛，内容新颖，条理清晰，组织合理。

高凯老师是我多年的朋友，我们都在大数据搜索与挖掘方向上从事教学、科研与开发工作。高凯博士严谨的治学态度、理论联系实际的做法以及敬业的态度也一直为我所学习。非常荣幸能够有这个机会来为高老师的新著作序，认真拜读后，我以为本书实战性很强，是大数据搜索与挖掘所需的上乘之作，是大数据“知著、见微、晓意”的必备工具，值得推荐！

2015.5.5

（张华平博士，副教授，北京理工大学大数据搜索挖掘实验室主任，ICTCLAS 及 NLPIR 分词软件发明者）

前言

Foreword

建立在分布式系统之上的大数据搜索与挖掘应用，是当今IT业的研究与工程实践热点之一。在DB-Engines公布的2015年度最受欢迎的数据库系统中，Elasticsearch名列前茅。作为开源分布式检索与数据处理平台，Elasticsearch不仅仅是一个数据库，它还是一个基于Lucene构建的开源、分布式、RESTful信息检索框架。基于Elasticsearch+Logstash+Kibana的信息处理架构，为编程人员提供了一种分布式可扩展的信息存储和全文检索机制以及基于Logstash的日志处理机制、基于Kibana的挖掘结果可视化机制。它不仅能对海量规模的数据完成分布式索引与检索，还能提供数据聚合分析和可视化。因此，从实战的角度掌握Elasticsearch、Logstash、Kibana的基本使用方法和技巧，很有必要。

大数据这个术语的出现，大概可追溯到基于Lucene的Apache开源项目Nutch。从2009年开始，大数据开始成为互联网行业的流行词汇，也吸引了越来越多的关注。物联网、云计算、移动互联网、手机与平板电脑、PC以及遍布各个角落的各种各样的传感器，无一不是大数据的来源方或承载方。可以说，大数据就在我们身边。从阿里巴巴、1号店、京东商城等电子商务数据，到QQ等即时聊天内容，再到Google、Bing、百度，又到社会网络与微博、微信等，都在生产、承载着大数据。随着信息处理量的增大，对大数据的分布式存储、快速搜索与挖掘显得特别必要。例如，挖掘用户的行为习惯和喜好，从凌乱纷繁的大数据背后找到符合用户兴趣和习惯的产品和服务，并对产品和服务进行有针对性的调整和优化，本身就蕴含着巨大的商机。但是，传统的基于关系型数据库管理系统的方法，在高效处理大数据时显得有些力不从心。虽然开源的全文检索工具Lucene能处理非结构化和半结构化的信息，但其某些版本在分布式处理方面的不足限制了它在大数据方面的应用。我们希望找到一个快速的分布式信息检索解决方案，希望它是一个零配置和易于上手的全文检索模式，希望它能够简单地使用JSON通过HTTP索引数据，更希望它支持分布式处理并支持系统扩展，能够实时搜索，并且稳定、可靠。

Elasticsearch 是一个基于 Lucene 的开源分布式信息检索架构和全文搜索工具。构建在 Elasticsearch 基础上的日志处理工具 Logstash 和信息可视化组件 Kibana，能有效衔接并高效处理由 Elasticsearch 索引的分布式数据，三者优势互补，各司其职，共同完成网络大数据分布式存储、倒排索引、全文检索、Web 日志处理、挖掘结果可视化这一整套的信息处理流程。目前，国内这方面的资料很少，仅有的几部译著所提及的 Elasticsearch 版本较低，且没有任何有关 Logstash 和 Kibana 的书籍。因此，我们萌发了一个想法，将 Elasticsearch、Logstash、Kibana(统称为 ELK)联袂奉献给广大软件开发者，帮助他们尽快熟悉 ELK 架构，并构建自己的 Web 应用程序，完成对分布式信息的检索与分析工作。

本书强调实践，内容新颖，条理清晰，组织合理。通过实战讲解的方式，让读者更好地了解 ELK 架构的实现细节。全书内容涵盖 ELK 简介、文档索引与处理、信息检索与过滤、信息统计与分析、基于 Java 客户端的 Elasticsearch 功能实现、Elasticsearch 配置与管理、基于 Logstash 的网络日志处理、基于 Kibana 的分析结果可视化、应用实例等多个部分。

全书由高凯提出写作大纲，第 1 章和第 6 章由高凯撰写并完成全书通稿和审校工作，其余各章均由高莘撰写，其中，第 1 章概述 Elasticsearch、Logstash、Kibana 的主要功能，对涉及到的一些概念进行简介，并从实用的角度出发，通过对实例的讲解，介绍索引、检索的实现机制；第 2 章对 Elasticsearch 中的索引、映射等进行说明；第 3 章介绍 Elasticsearch 中的检索功能；第 4 章介绍基于 Facets、Aggregations 的数据聚合与统计功能；第 5 章从工程实践的角度，介绍面向 Java 客户端的 Elasticsearch 部分功能的设计与实现；第 6 章介绍 Elasticsearch 的配置及一些高级功能、监控等的使用；第 7 章介绍日志处理及 Logstash 的应用；第 8 章介绍基于 Kibana 的可视化技术；第 9 章给出一个综合应用实例，该实例从网页采集、处理、存储、索引、日志处理、可视化展示等入手，介绍了基于 ELK 的分布式信息检索与日志挖掘解决方案。

本书的顺利完成也得益于参阅了大量的相关工作及研究成果，部分内容源自 Elasticsearch、Logstash、Kibana 的官方文档。在写作过程中，也参考了相关文献和互联网上众多热心网友提供的素材，在此谨向这些文献的作者、热心网友以及为本书提供帮助的老师，特别是那些由于篇幅所限未及在参考文献中提及的相关文献的作者和网站，致以诚挚的谢意和崇高的敬意。

由于我们的学识、水平均有限，书中不妥之处在所难免，恳请广大读者批评指正。

编　者

2015 年 5 月

Contents

目录

Chapter 1

第1章 概　述

> "Elasticsearch is a flexible and powerful open source, distributed, real-time search and analytics engine. Elasticsearch gives you the ability to move easily beyond simple full-text search.
>
> Logstash helps you take logs and other time based event data from any system and store it in a single place for additional transformation and processing. Logstash will scrub your logs and parse all data sources into an easy to read JSON format.
>
> Kibana is Elasticsearch's data visualization engine, allowing you to natively interact with all your data in Elasticsearch via custom dashboards."
>
> ——http://www.elasticsearch.org/overview/

随着大数据、大型综合网站以及 Web 2.0 技术的普及，越来越多的软件开发者需处理海量异构信息的索引、检索、日志挖掘、可视化等和信息检索与大数据挖掘相关的业务。虽然 Lucene 是许多互联网公司的标准信息检索工具，但它无法在一个合理的时间内存储和检索海量的大数据，不具备良好的可扩展性，一般也不适合分布式大数据搜索、挖掘和云计算环境。而在 DB-Engines 公布的 2015 年 5 月份最受欢迎的数据库系统（含 NoSQL 数据库）中，Elasticsearch 名列前茅[Db-engines, 2015]。

作为开源分布式搜索与数据处理平台，Elasticsearch 不仅仅是一个数据库，它还是一个基于 Lucene 构建的开源、分布式、RESTful 信息检索框架，能够实时搜索，并且稳定、可靠，使用方便，支持通过 HTTP 使用 JSON 进行数据索引。基于 ELK（注：本书中的 ELK 是指 Elasticsearch＋Logstash＋Kibana，后文同）的架构为编程人员提供了一个分布式的可扩展的信息存储和基于 Lucene 的信息检索机制、基于 Logstash 的日志处理机制、基于 Kibana 的挖掘结果可视化的机制。在一个典型的使用场景，一般用 Elasticsearch 作为后台数据的分布式存储和全文检索，Kibana 用来前端的可视化展示，Logstash 在其过程中担任相关日志加工和"搬运工"的角色。ELK 架构为数据分布式存储、可视化查询和日志解析创建了一个功能强大的管道链。三者互相配合，取长补短，共同完成分布式大数据处理工作。

首先，Elasticsearch 是一个开源的分布式信息检索框架，具备高可靠性，它提供多种管理工具，各种相关插件也可方便地集成到 Elasticsearch 中。它对外提供一系列基于 Java 和 HTTP 的 API，可用于分布式索引、检索、日志分析与数据挖掘等，且大多数配置是可以修改的。因此，很多国际知名企业都在使用 Elasticsearch 完成分布式数据处理工作。例如，Github 已升级了其代码搜索程序，并将核心架构由 Solr 转向 Elasticsearch；Wikimedia 也启用了由 Elasticsearch 为基础的全新搜索框架。

其次，Logstash 可以对相关的网络日志进行收集、分析、转换等处理工作，并将其存储在 Elasticsearch 供以后使用。其实，Logstash 本身并不产生日志，它仅仅是一个可接收多种多样的日志输入、经处理后转发到多个不同目的地的“管道”。最后，Kibana 可以帮助可视化数据日志，并提供友好的可视化界面。

学习 ELK，对于大数据处理、信息检索及搜索引擎研发、日志处理与分析、挖掘信息可视化等，对于设计高效的大型商业网站，都具有重要的现实意义。本书主要介绍 Elasticsearch 分布式信息存储与检索解决方案，并结合 Logstash、Kibana，介绍面向大数据搜索与挖掘的处理方法。作为全书的“引子”，本章介绍 Elasticsearch 的背景和简单使用，并通过一个例子介绍 Elasticsearch 的索引、检索流程和实现方法。有关 Logstash 和 Kibana 的相关内容，在本书后续章节中进行介绍。

1.1 Elasticsearch 的安装与简单配置

“工欲善其事，必先利其器。”要想了解 Elasticsearch，要从该软件的安装入手。Elasticsearch 的安装非常简单，几乎是“开箱即用”的。当然，前提是需要先下载 JDK，并配置相应的环境变量，同时确保系统可用内存大于 2GB。

> Tips：建议使用 JDK7 或 JDK8，低版本的 JDK 会对 Elasticsearch 的使用造成不利影响；建议设置 JAVA_HOME 环境变量。

下面对 Elasticsearch 的安装进行说明。进入 Elasticsearch 官网 http://www.elasticsearch.org，找到对应的 Elasticsearch 软件版本下载。如果是在 Windows 系统下使用，可以下载 ZIP 格式的安装包。对于工程开发人员来说，往往需要在 Elasticsearch 软件中集成一些其他插件和工具等。因此，建议初学者可以首先从 Elasticsearch 的 RTF 版本入手。可以到 https://github.com/medcl/elasticsearch-rtf 去下载 Elasticsearch 的 RTF 版本。

> Tips：RTF 是 Ready To Fly 的缩写，这是一个集成了基本插件（如服务封装、中文分词、mapper-attachments、transport-thrift、tools.carrot2 等插件）的并带有示例程序的可直接上手的简易工程版本。

解压后会看到其目录结构。Elasticsearch 包含的主要文件夹及功能如下(以 RTF 版为例):

- bin——含有运行 Elasticsearch 实例和管理插件的一些脚本。
- config——主要是一些设置文件,如:elasticsearch. yml 和 logging. yml 等。对 elasticsearch. yml 和 logging. yml 文件中相关配置的说明,可参阅本书后续章节。
- lib——包含一些相关的包文件等。
- plugins——包含相关的插件文件等。
- logs——日志文件。
- data——Elasticsearch 中存放数据的地方。
- works——临时文件。

Tips:如果 Elasticsearch 运行在专用服务器上,一般经验是分配一定的内存给 Elasticsearch,可以通过修改 ES_HEAP_SIZE 环境变量来改变这个设定,它控制堆大小。在启动 Elasticsearch 之前应该把这个变量改到预期值。关于这个数据的设置可参见相关手册。一般来说,如果在日志文件中发现带有 OutOfMemoryError 错误的输出记录,则应考虑将环境变量 ES_HEAP_SIZE 取值调大,建议该值不应超过总可用物理内存的 50%(剩余内存可用作高速缓存,提高检索性能),这样可以极大地提高搜索性能[Rafa,2015]。

可选择 Elasticsearch 使用的中文分词器。打开 config / elasticsearch. yml 文件(注:yml 是一种简单的数据描述语言,语法比 XML 简单,适合用来表达或编辑数据结构,并完成各种设定等)。在这个文件中的 index. analysis. analyzer. default. type 部分指定使用的中文分词器。代码段 1. 1 是选择使用 IK 分词器并对其进行设置(注:在 Elasticsearch、Logstash、Kibana 等相关的配置文件中用 # 表示注释信息。为便于统一表述形式,本书多用//表示注释说明信息,仅在第 7 章有时用 # 表示注释说明信息)。

```
//代码段 1.1: 在 elasticsearch.yml 中设置中文分词算法
index:
  analysis:
    analyzer:
      ik:
        alias:
        -ik_analyzer
        type: org.elasticsearch.index.analysis.IkAnalyzerProvider
      ik_max_word:
        type: ik
        use_smart: false
      ik_smart:
        type: ik
        use_smart: true
index.analysis.analyzer.ik.type: ik
index.analysis.analyzer.default.type: ik
```

Tips：在文本被索引前需要经过分词处理，这项工作一般由 Analyzer 类完成。Analyzer 类是个抽象类，对应不同语言的文本，应该从 Analyzer 派生出特定的 Analyzer。IK Analyzer 是一个开源的基于 Java 语言开发的轻量级的中文分词工具包。最初它是以开源项目 Lucene 为应用主体的、结合词典分词和文法分析算法的中文分词组件。从 IK Analyzer 3.0 起，发展为面向 Java 的公用分词组件，且独立于 Lucene 项目。另外，yml 配置文件多数内容是被 # 注释起来的，需要修改某部分时，可以删掉注释标记并进行相关的配置即可。yml 内容要求严格执行规定的缩进格式。

进入 Elasticsearch 的 bin 文件夹，运行 elasticsearch.bat 文件，启动 Elasticsearch。

Tips：若关闭 Elasticsearch，可在 shell 环境输入命令（Elasticsearch 默认占用 9200 端口）：

```
curl -xPOST http://localhost:9200/_cluster/nodes/_shutdown
```

之后，打开浏览器，输入 http://locaohost:9200，会显示类似图 1.1 的内容。其中：

- status——发出请求后的 HTTP 的状态代码，显示“200”表示正常。
- name——Elasticsearch 实例的名字，默认情况下它将从名字列表中随机选择一个。其设置同样是在 config / elasticsearch.yml 文件中。
- version——版本号，以 JSON 格式表示了一组信息，其中的 number 字段代表了当前运行 Elasticsearch 的版本号，build_snapshot 字段代表了当前运行的版本是否是从源代码构建而来，lucene_version 表示 Elasticsearch 所基于的 Lucene 的版本（图 1.1 显示该版本是基于 Lucene4.10.2 而构建的）。
- tagline——包含了 Elasticsearch 的第一个 tagline： "You Know, for Search."。

图 1.1 Elasticsearch 启动后的界面

图 1.1 中出现了 JSON 格式的数据。JSON（JavaScript Object Notation）是基于 Javascript 的轻量级的数据交换格式，是独立于语言的文本格式。在 Javascript 中，处理 JSON 数据不需要任何特殊的 API 或工具包。利用 JSON 可简单地表示半结构化数据，而